When is it Justifiable to Ascribe Mental States to Non-Human Systems?

Thesis

When is it justifiable to ascribe mental states to non-human systems?

Susan Alice Jane Stuart

Submitted for examination for the degree of PhD, from the Human Cognition Research Laboratory,

1 November 1992

DATE OF SUBMISSION : 5th NOVEMBER 1992
DATE OF AWARD : 25th JANUARY 1993

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

In this thesis I shall attempt to show when it is, and when it is not, justifiable to ascribe mental states, of the type that we associate with the complex cognitive behaviour of human beings, to non-human systems. To do this I will first attempt to give a fundamental explication of some of the problems that underlie our ascription of mental states to other human beings, non-human animals and machines, after which I will tackle the problem of whether or not any ascription of mentality can ever be completely vindicated.

Then I will look at the issues of complexity and the distinctions that hold between the capabilities of various systems, both natural and artificial. The result of this will be a more comprehensive understanding of what characteristics are necessary for the possession of such capabilities. I will go on to argue that a positive relation exists between a system's architecture and its capability to behave or act in ways that can be classed as one of a number of mental states such as 'knowing', 'understanding' or 'believing'.

I shall look at the ways in which machine states and mental states have been examined using hierarchical stratifications for these can offer us some indication of the correlation that exists between simple systems and the low level actions of which they are capable, and the more sophisticated actions of which only progressively more complex systems are capable. However, I shall put forward arguments to demonstrate that this is a feasible strategy when dealing with the innards of a machine but not for dealing with the innards of the mind.

Throughout the thesis I shall try to clarify the inexplicit or clouded notions of subjectivity and intentionality, for one of my aims is to demonstrate that the notions of subjectivity and awareness are more important than intentionality in the distinction between human and non-human systems.

Acknowledgements

Many people deserve my gratitude for the support and encouragement they have shown me. For this work to have reached the stage it has is due in no small way to my good friend Arthur Stutt who has shown great patience and perseverence when working with me; I am indebted to you, Arthur. I would also like to thank Professor Bill Lyons for reading and commenting on the pieces of work that I sent him, and for always being supportive. Tony Hasemer has my gratitude for being there when I felt like giving in and persuading me to continue to struggle against what often seemed like an insuperable task. And I would also like to thank George Kiss for his initiative and advice, even when we did not always see eye to eye.

I have no end of admiration for my friends Clive, Stuart, Mike and Paul who were able to make me laugh in my final year, and for keeping me talking and believing in what I was doing; for this I wish you all much success and happiness. A special 'thank you' must go to Clive for his counselling on many a tearful afternoon. Thanks also to Anne Carson for putting at my disposal her tremendous organisational abilities and general good-heartedness, and to Danny Waite for being a kind and reliable friend.

In general I would like to thank everyone else at the Human Cognition Research Laboratory for making it a friendly environment in which to pursue my goals. I am grateful with all my heart to the Open University for awarding me the scholarship with which I was able to carry out this work.

Outside HCRL I would like to thank Shirley Coulson for being a good listener and a dependable friend, and David Broadhurst for giving me much needed reassurance. But most of all I would like to thank Norman who has had faith in me when I have not, who has shown me that it is healthy to have confidence in oneself and who has continued to love me through even the most formidable conditions.

This thesis is for my mother, Jane, who has made all things possible and whom time has given me the good fortune to know.

Contents

Table of illustrations

Preface

I will say here a few words in defence of my idiosyncratic use of certain words that will appear frequently in the following thesis.

To begin, I use 'organic' and 'inorganic' to distinguish between living or natural entities and those that are made artificially. My decision to do this is largely based on the amount of unnecessary trouble that I have seen other writers create for themselves when using 'machine' to cover all types of system. I think that even when 'machine' is used to describe biological systems, as Searle uses it, there is still the underlying connotation of something mechanical which is at best unhelpful if we are to maintain any distinction.

The second defence I will make concerns my use of the term 'system'. I use 'system' to describe any process that is not necessarily a living and breathing entity but which can exhibit similar behaviour; that is, to cover both organic and inorganic entities. For instance, in the work that follows I will be arguing that there are some machines, such as thermostats, that can react to changes in their environment, yet there are some lower order animals, such as sea-cucumbers, that behave in a similar manner; I will use the generic word 'system' to describe both types of entity.

A lexicon

Analogue - A term adopted by Dretske to describe any and all informational input before any selection has been made and processing carried out.

Analytic - By 'analytic' is meant that the concept of the predicate is contained in the concept of the subject as analysis of the terms would disclose.

Asymptote - A line that continually approaches a curve but never actually meets it. **Asymptotic** is the adjectival form.

Command-line - An instruction typed in by the user, usually in a formal language, to direct the computer.

Conceptualisation - Dretske's term for the analysis of incoming information to form a concept from which knowledge is attained and beliefs formed about the world.

Contingent - Something that is 'contingent' may exist and also not exist, which is to say that for its existence it is empirically dependent upon the world being in a particular state at a particular time.

Database- A corpus of information stored in a computer which can be processed by the computer and information retrieved when required.

Digital - Another of Dretske's terms, this time referring to the focusing in on one specific object or event in the visual field from which the semantic content is then reached and extracted by the process of **'digitalisation'**.

Nomic - A term adopted by Dretske to mean that which is dependent upon empirical laws that hold in the world.

Transition Graph - Most closely resembling a flowchart, consisting of labelled circles that represent 'states' and a series of arrowed lines that either loop or go on to another 'state' or circle. The input state is indicated by an input arrow and the final state by two concentric rings.

1. Setting the scene

1.1 An introduction

In setting the scene I will attempt to justify my work in three of the most pertinent areas: philosophy of mind, psychology/cognitive science and artificial intelligence. The issues that I shall be dealing with are not just contemporary ones, but matters about which there has been a great deal of controversy and debate for many decades.

In the philosophical areas of epistemology and the philosophy of mind scholars have been absorbed for centuries in disputation about the mind/body problem. A distinction between mind and body has been posited, disputed, withdrawn but never finally settled. It is a distinction that has found its way into Artificial Intelligence (AI) in the form of how physical systems can be described using mentalistic terms.

Specifically my concern is with the nature of the internal states of the system. Both the 'carbonists' and the 'siliconists' agree that mental properties exist and that they are the properties of physical systems. The distinction between carbonism and siliconism is simply that the former believe that only organic systems can possess mental properties, whilst the latter believe that these types of properties can be instantiated in mechanical systems. For the carbonist it is the fact that the organic system is composed of carbon molecules that is sufficient to set it apart from inorganic or mechanical systems. On the other hand, siliconism, or functionalism, states that anything that exhibits the appropriate 'understanding' behaviour can be said to 'know' what is happening, and in this way there can be no real distinguishing features between organic and inorganic systems.

One of the main reasons for embarking on this work is that in AI there have been many attempts to simulate or reproduce mental states, whether in the von Neumann machines or the neural networks of connectionism; the positions have been staked out, weapons raised in anger, but the war has only just begun. Whatever position we decide

to adopt, any simulation of mental states will be sadly lacking if we have only an incoherent or 'half-baked' idea of what 'mental life' is all about.

Because our picture of the brain or mind is incomplete it follows that our models will be lacking in some element, or elements, that are essential. It is this area, this essential element, that marks the difference between the simulation and the duplication of a mental state, which will be the subject of my work. What follows is intended to reduce the fuzziness that is notoriously associated with this area.

1.2. Mental life, mental states and intentional behaviour

In this section I shall offer a brief explanation of the central ideas that underlie the main body of the thesis. I will also set out the method that I intend to follow and what detailed work will be entailed in the attempt to reach my final proposed end.

One of the main objectives that first needs to be resolved is to create some sort of holistic view of what 'mental life' is considered to be. This is of the utmost importance since it is 'mental life' and in particular the mental action of organic systems that will count for at least half of the subject matter of the thesis.

One of the main problems with mental life is figuring out just what sorts of things go to make up the mental aspect of the system that is to be examined. Many things can be subsumed under the vague title of mental life; and most commonly we think of the ability to recall past events, having perceptual skills, solving problems, having ideas and being able to entertain abstract notions. For our purposes now we will be concentrating on particular mental occurrences within each of these vast areas. These occurrences have, perhaps somewhat unhelpfully, been described as 'mental states'. I say 'unhelpfully' because when the word 'state' is used there is a tendency to think of some sort of mode of existence, and, as I shall now explain the 'existence' that mental states manifest is a most unusual one.

To define the characteristics of a mental state is probably easier if I first say what it is not. Foremost, mental states are not objects or ostensible states of affairs; however, for their manifestation they do require a world of objects and events.

There are two definitions that I will look at in detail. The first is of mental states in terms of 'experience'. I am looking for a less interactive term than *mental attitudes* because the word 'attitudes' suggests that the system has processed its information to the extent of having formed an opinion about the world and itself in relation to that world and to have a mental attitude is to have a more active interaction with the world than I wish to demonstrate here.

What I want to claim is that it is possible to have mental experience without necessarily having any self-reference in that experience. For example, the experience had by X when it stands in a relation Y to events Z. So to have a mental experience of something simply requires that the system is in a set of circumstances, for example, where the weather is changeable and John has to venture outside then he might without much thought pick up an umbrella before he leaves his house. Quite simply his experience of looking out the window and observing inclement conditions urges him to carry protection against the elements. It was not necessary for him to go through all the mental states connected with that particular state of affairs. This kind of notionally interactive experience might be equated to running on 'auto pilot'.

The second definition I will turn to is one offered by Myles Brand in *'Intending and Acting'* that states that all mental states or 'events' are characterised by mental attitudes for which there is some object. As he states "I take a mental attitude to be a mental event that has an object.....The objects of mental attitudes, I will argue, are properties....That is, all attitudes....can be analyzed in terms of attitudes that take properties as their objects."[1] All mental attitudes are reflexive, that is, all propositional attitudes, like "intendings are self-referential".[2]

I will be arguing in a similar vein that a mental state becomes a mental state when a mentally active system, that is, one that is assumed to possess mental life, perceives itself to be in relation to events that are external to it. In the language adopted by Brand a person attributes a property to something else when he attributes to himself the complex property of standing in a unique relation to that person or thing which has that property.

If mental states are the reflexive relation between the system and its environment, and the environment is continually changing, then the mental states possessed by the system must respond at a similar rate if it is to deal with the change successfully. This notion of changing mental states can be helped by further defining 'state' as that which is 'affected'. Being in a relation to changing events in the external world will *affect* the internal mental states and cause them to change. But it must be borne in mind that this is only one possible definition of mental state that uses reflexiveness as a property.

The agent or perceiver is the system with the mental state and the mental state can be of the form 'believes that', 'hopes that', 'longs for' or any similar phrase that describes a mental state. Terms such as these are classed as 'intentional' which means that they describe the intentional relation of an agent to some particular state of affairs. In Chapter 2, which is a critical review of some of the most pertinent literature, I will be examining intentionality in relation to a number of influential writers.

Intentionality is certainly one of the most significant mental properties that indicates a link between the system and its environment. With the action that is consequent on intentional thought it can be assumed that the system has interpreted the information that it has coming in. This interpretation is called 'processing'. Processing of information and the occasion of action suggests that the system has had to do something more than just process the information. It has moved through a perceptual phase, an information processing stage, and on to some selected course of action.

1.2.1. Understanding and knowing - epistemic states

The selection of an action requires that the system be capable of manifesting some state that might be called 'epistemic'. It is a state that has been reached by having processed the information, 'understood' it to some degree and, on the basis of this 'knowledge', either acted or formed a belief and used that new belief to reform its framework of interpretation for future use.

This is a complex procedure and the problems being addressed are difficult ones that would each by itself merit extensive discussion. This will be carried out, first in

brief, in sections of the literature review as an overall look at what other theorists have concluded count as mental states, and then in more detail throughout the third chapter with an examination of the notion of the ascription of mental states, and specifically 'intentionality', to other human beings and other non-human systems.

The opening sections of chapter two examine mental states, mental acts and intentionality. This is followed by an examination of understanding as an instance of intentionality in relation to the works of Searle, and in particular with reference to the Chinese Room argument, after which I will take a more general look at what behaviour has to do with the possession of certain types of mental properties by a system.

Quite simply what Searle says is that it is not possible to understand symbols by virtue of their being symbols. In his Chinese Room Searle receives Chinese symbols but because he has no knowledge of Chinese he is unable to understand the symbols. He has, however, the capacity to match symbols to other symbols in the book and hand the matched symbols out of the room. Searle argues that this is not understanding even though there is still all the associated understanding behaviour: the symbols are matched correctly with other symbols and the appropriate response is elicited from the room. "A program merely manipulates symbols, whereas a brain attaches meaning to them."[3]

I agree with Searle, and I would also wish to argue that the exhibition of intentionality, in the form of understanding behaviour, is not a clear indication that the system 'knows' what is happening in its environment. I shall set out arguments to demonstrate this. Briefly, then, what I am saying is that it is not possible to say what properties of the class of intentional behaviour are made known to us by that behaviour alone.

A closer examination of the possession of mental properties in relation to the exhibition of certain kinds of action might reveal that because it is not possible for us to look directly at mental states or mental attitudes we are left with the fact that the behaviour of a system is the only real indication of what properties a system might possess. As we shall see from the discussion of the recent work of Stanley Rosenschein that although the type of 'knowing' he refers to is very basic it is still

suggestive of the inorganic system, being in possession of 'primitive epistemic states'. I shall assess the validity of this claim in relation to other similar claim.

As mentioned above the functionalist response to my claim is that the right sort of 'understanding' behaviour is all that is necessary for the machine to be said to understand. Using this tactic the ascription of epistemic states to inorganic systems ceases to be a problem and the success or failure of this strategy will be assessed critically. Involved in this will be an examination of the proposal that these terms have a metaphorical use in AI due to an over-extended use from organic to mechanical systems. If this is the case they have no place in the literature of AI that deals with inorganic systems.

1.2.2. Selectivity

Moving on from this another significant topic is the role that 'selection' plays in what a system is capable of doing. For instance, systems vary quite substantially from those that have no selective capabilities and which inhabit a fixed environment to those that can select the events in their environment that they will attend to and those that they will choose to ignore. To have selective capabilities of this sort the system has to be aware that it is part of a continually changing environment.

To be able to select these events the system must be able to see itself in relation to those events and the possible advantages or disadvantages they will have for it. There must be some way in which the system can assign priorities to the information it receives and respond to it in the most appropriate manner that will maximise its own chances of survival.

1.2.3. Subjectivity

Another significant aspect of mental life is subjectivity, and two things come to mind that are important when first considering it. Firstly the subjective nature of the interpretation of incoming information; and secondly the subjective aspect that necessarily accompanies the action of the system. Once the role of subjectivity has been further clarified in relation to mental life it ought to be possible to evaluate to what

extent it is necessary for the adoption of an intentional stance towards objects in the environment. It should also be possible to assess whether subjectivity is a necessary criterion for the possession of mental states, or whether it is a matter of the complexity of a particular type of mental state. For example, "knowing" might be a simpler mental state than "believing", and it might be that it is only when encountering the issue of belief, that the ability to interpret information subjectively becomes necessary before it is possible to move from "knowing that 'x'" to "believing that 'x'".

A problem that arises directly from this is whether or not the action of a system is a reliable source for the ascription of a particular mental state. There seems to be a leap of some kind between the action that takes place and the ascription of the 'state of mind' the system was in when the action took place. It is hard to assess what this missing link could be except perhaps some sort of interpretive process.

When addressing this question there are several other things that will have to be taken into account, such as whether or not the system is complex enough to occupy such a state or to be capable of deception, that is, that it might behave one way whilst concealing that it is occupying an inconsistent mental state. The behaviour is inconsistent with the mental state or intention that the system has - unless, of course, the systems intention was solely to deceive.

The issue of what behaviour is most advantageous to the system, and that some complex systems are capable of selecting the course of action that is most suitable to them, is an issue for both subjectivity and self-reference. In one sense we could be said to have come full circle to an indication of the advantages that subjectivity can offer a system.

I shall be arguing that subjectivity must exist for some reason and I shall examine the advantages that it offers the organic system over the mechanical one. These advantages can be both ecological and evolutionary. The emphasis at this stage will be on what in particular the complex system is offered by being in possession of a high-level of awareness that I will equate with self-consciousness.

There has been a great deal of discussion about self-consciousness but very little concise and quotable work has yet been produced. In my thesis I prefer to think of consciousness, and self-consciousness in particular, not as an 'on/off' switch, that is applied to organic entities such as frogs, cats and human beings but not to televisions, plants and tea-cups, and more like something that can be applied to any system that occupies a higher level awareness. This requires that I clarify the notion of awareness, which will then permit me to relate different degrees of awareness to a range of functions that can be carried out by a system.

If we accept a notion of intentionality as being attributable to a variety of systems, and we realise the importance of subjectivity in the actions and potential actions of the system, then the most natural assumption to make is that the distinction between organic and inorganic systems resides in the fact that some organic systems are capable of possessing a high-level of intentionality, a high-level of awareness, that is self-consciousness, and a subjectivity with which it has the flexibility to respond to changes in its environment.

I will be arguing that subjectivity, in a way similar to self-consciousness, is a high level capability that requires not only that the system can be self-referencing, but also that it is self-aware, which when examined more closely may turn out to be a more complex notion than just being capable of reflexiveness. A related ability of which I will maintain a subjective system needs to be capable is the creation of symbols, the arbitrary assignment of meaning to those symbols and the use of those symbols in shared communication.

1.2.4. Intrinsic meaning

The arbitrary assignment of meaning to symbols is a much more complicated area that I will begin to look at by analysing what is entailed in the notion of 'intrinsicality' of meaning. The latter deals with the construction of symbols and whether their meaning is attributed by the designer of the system or simply intrinsic to them. In a more detailed argument I will argue against Harnad's claim that meaning can be

grounded in the sense data or sensory information of the symbol; and as a starting point my first argument will be that common sense alone suggests that the meaning of a symbol is something that is attributed to the symbol system since it will be constructed by a reflexive organism that uses symbols.[4] In brief my second argument against intrinsic meaning is that symbols are constructs and the semantics of a construct cannot be 'made' intrinsic; something is either intrinsic or it is not. If it is not intrinsic it is, per se, attributed. I shall also refer quite closely to Rosenschein's work on 'interpreted symbolic systems'. His work states that the interpretation of a program is dependent upon the designer attributing it; so that without the programmer the program would have no meaning.

1.3. An interim summary - "The central ideas"

To recapitulate, I have explained the need for more work in the area of possible mental properties of inorganic systems, and I have opened up the debate so that we can now look forward to the notions of philosophy, cognitive psychology and the recent work in AI being placed in an academic setting.

The next step is to ask what information is necessary for a more complete picture of mental life, since it is 'mental life' that AI, in particular, is interested in. The first prerequisite was that I take a careful look at what constitutes mental life. The conclusion was 'mental states' and the enquiry was then developed to an examination of the properties of those states. The most obvious next step was to define such states and then to relate them to the observable behaviour of the system.

In an attempt to do this successfully it is necessary to consider the organic system in totality, that is, to look at the importance of subjectivity in relation to two things; the first was how it sees itself in the environment it occupies with which it interacts, and the second was how being subjective allows the system to create and assign a meaning to the symbols it is to use. The aim is to show that it is a definite advantage for a system to be subjective, and that for complex cognition, that is, the sort of cognition we associate with the human system, a system also needs to be self-conscious, flexible

enough to select appropriate information, and create and assign a meaning to symbols, and then use those symbols to communicate in the form of a shared language. This last requirement is arguably the most important for it is only through the possession of it that any system would be at all capable of expressing its self-conscious capabilities and the subjectivity of its judgements.

1.4. Outline of the thesis

In this section I will set out briefly the procedure that I will follow chapter by chapter. By this stage I hope that I have made clear the nature of the work that will follow and the conclusions at which I will be aiming to arrive.

The second chapter will be a review of the literature that is covered in this thesis and related texts that have been useful, if in some cases, only indirectly. As mentioned above the main concern of the review will be the nature of intentionality, the problems that surround it and its relation to the work that has been done with regard to other mental attitudes and properties.

In chapter three of the thesis I put forward the claim that the ascription of mental states to other systems is done by analogy with our own mental states. The only mental states of which we have any truly direct experience are our own and such experience is always subjective. No one but me can experience my mental states and I cannot directly experience those of anyone else except vicariously through their descriptions of them. Thus it is that the only states which any system experiences directly are its own states.

It is possible to ascribe mental states to other systems either linguistically or behaviourally. By linguistic ascription I mean using propositional attitude statements to describe X's behaviour, for example, that 'X believes Y' or that 'X fears Y'. When ascribing mental states to another human being the person to whom the states are ascribed can offer corroboration or denial of the ascription by themselves using propositional attitude statements. The behavioural ascription of mental states takes place, often without the awareness of the ascribing system, when that system perceives that 'X' possesses a set of internal characteristics. It is a sort of 'poking and fiddling'

approach, a brute physical enquiry by which we become aware of the internal state of the system.

One example that highlights differences in behavioural ascription is my different attitudes to a thermostat and a video recorder. By poking at a thermostat and looking inside it I will realise from its simple design that it has only a limited number of functions, but I will ascribe a much more complex set of behaviours to a video recorder, with all its buttons and its complex array of internal wiring, than I would a simple thermostat. An example that concerns organic, but non-linguistic systems, would be that I might see a snarling dog at the end of a street that I wish to walk along and my apprehensive behaviour would act as a tacit ascription of anger to the dog and fear on my behalf. In both examples nothing has been expressed using language, and our behaviour is only based on what we perceive the state of the other system to be.

We usually ascribe mental states on the basis of consistent human-like behaviour, which is to say behaviour that is consistent with how we imagine that we would react were we in the context of the behaving system. (Because my states are the only ones to which I have access and they are examples of human mental states then the behaviour has to conform to my human behaviour.) Thus any system that behaves 'as-though' it possesses human mental states is often considered to have such states.

Why do we ascribe mental states to systems other than ourselves in the first place? Well we spend a great deal of time interacting with inorganic systems that can perform mental-like tasks, and ascribing mental states to them is a useful predictive tool that facilitates interaction and communication between them and us; that is, between human beings and what are perceived to be 'intelligent' systems. So that even if the system, whilst exhibiting signs of mentality, is still known to be inorganic, it is probably best, or at the very least useful, to behave towards it as one would towards a human being that is known to have a brain and a complex mental life.

1.4.1. How does the issue of complexity relate to this?

In chapter four I relate complexity to the problem of ascription in three different ways; the first is that a system has to be of a fairly high degree of complexity for it to be capable of acting in a way that could 'persuade' us that it is sufficiently 'human-like' to be ascribed mental states; secondly, the process of ascription itself is complex one, whether it is being done linguistically or behaviourally; and thirdly, behaviour is a complex relation of architecture and environment, so that the internal design of the system and its environment afford the system a variety of capabilities, some of which are complex and others not so complex.

A subsidiary element of the first aspect of complexity is that if the ascription is made linguistically it must be made by systems that are capable of using language to form and express propositional attitude statements. Systems such as this must be 'symbolic' in the sense that they are capable of creating abstract symbols and subsequently assigning meaning to those symbols. I propose that the semantic interpretation of a symbol system cannot be made intrinsic to that system, and it follows from this that the meanings that the symbols possess depend entirely upon the choices made by the designer or programmer who is assigning their meaning.

1.4.2. The relationship between a system's architecture and its capabilities

In chapter five using two already established examples I shall put forward evidence to show that a relationship holds between the capabilities that a system has [which are a function of its mental states] and its internal design or architecture. The first is a hierarchy that Chomsky constructed in 1959 to compare the variation in structure of four different machines with their related capabilities. One of the limitations of this work is that it only deals with machine states. In a more recent work Dretske (1981) does the same sort of thing, but this time with mental states. Both hierarchies are limited in their own ways; Chomsky's because it does not show that the same hierarchy can hold in the case of organic entities with mental states, and Dretske because he does not complete his hierarchy by suggesting systems that have capabilities that are

comparable to his second level of intentionality. These limitations offer me the chance to produce a fuller description that relates both to organic and inorganic systems and also to mental and machine states.

Towards the end of the chapter I offer an explanation for why hierarchies such as Dretske's are bound to fail. For example, I argue that because we are dealing with fuzzy concepts, and I include all mental states in this category, it is mistaken to try to delineate them into discontinuous, discrete categories of one state or another. Chomsky's hierarchy is not in jeopardy in the same way because he is dealing with a straightforward set of machine states and tasks that can be described and set out in a finite number of discrete steps.

1.4.3. Advantages of cluster diagrams when examining mental states

Having argued in chapter five that hierarchies are not the most useful way of envisaging a relationship between a continuous set of mental states, I begin chapter six by proposing some alternative ways of exemplifying just such a set. On offer are an assortment of cluster diagrams which might be used to express the overlapping nature of mental states and in which systems such states exist in some form or other. Diagrams of this sort are often used as taxonomic devices for decided the category of one species or another. One of the main points in this section is that no perfect set of axes exists within which we can define the nature of fuzzy or vague concepts; thus every graphical interpretation depends upon what is to be plotted and what it is to be measured against that appears on each axis.

Towards the end of the chapter I examine Sloman's most recent work which concentrates on design and the 'design space' in which different architectures occupy different points. Sloman argues that for a system to be capable of different activities it would need to occupy different points in the design space. Thus for a system to be capable of more complex things it needs a more complex design space. For Sloman the human being has a very rich and complex design space and from this it can be inferred that it has also a rich and complex repertoire of possible behaviours. But being rich and

complex is not sufficient for complex behaviour for in the design space we need also to look at what the system needs to sustain its existence in the environment it occupies. So how something is capable of doing what it does matters and not just that it can do it.

1.5. Conclusion

Bearing all of this in mind the thesis is concluded in chapter seven with a look at the advantages that the human system has over all other non-human systems. These advantages range over a great many things and I shall describe only the most significant in this present document. One advantage is that the human system is the only one that can create and arbitrarily assign meaning to a set of symbols. A second is that from the wealth of incoming perceptual information the human system has the flexibility to select the piece that is most appropriate to it whilst ignoring or storing other pieces for future use.

A third and very important advantage that the human system possesses is to be capable of the subjective interpretation of its incoming information. This subjectivity is personal and infinite in nature, whether that infinity is of the human being's potential environment or of its forms of communication that can either be linguistic, as we have seen in the creation and use of formal symbol systems, or non-linguistic in the form of body language. That human beings are subjective in their judgements is more significant than their having intentionality for even plants, thermostats and moles can behave intentionally in their own ways. The plant, for instance, is heliotropic, geotropic and hydrotropic, and its process of homeostasis is like a control centre that directs the plant to its sources of nourishment.

This chapter is brought to a close with an examination of what sort of architectural requirements are necessary for such a highly developed and complex cognitive system to exist and behave in the ways that it does, and how it would be best to show that a system's capabilities are related to its architectural complexity and the extent of its perceptual domain. In its graphical form it is finally possible to show that the human system is the only system that can possess a state of 'full blown' self-conscious

awareness, and that although lots of other systems can occupy states of varying levels of complexity, none, but the human, language using, system is capable of the full gamut of known mental states.

Endnotes:

[1] Brand, M (1984) *Intending and Acting*, MIT Press, p.85

[2] Ibid. p.92

[3] Searle, J. (January 1990) *Scientific American*, Vol 262, No.1, p. 20

[4] Here 'reflexive' is intended to mean the sort of system that can be self-referring.

2. Literature review

2.1. Introduction

This chapter consists of a review of some of the literature, from a huge library of work, in the area of intentionality, mentality and mental states. My overall aim will be to address the problem of what counts as mental life and I will begin by examining what we consider mental states, in their variety, to be. This will bring me to an investigation of the difference between mental and physical acts, and how mental actions, such as intentionality, can be expressed using propositional attitudes. From here I will introduce intentionality by examining the philosophical work of Brentano, which, although written in the last century, is still the subject of much inquiry today. A great many of his points, notably, the directedness toward the objects of intentional behaviour and the 'immanent objectivity' of the objects, the mind as a faculty of awareness and the human mind's capacity for reflexive awareness, have not been adequately resolved and these will be issues that arise throughout this chapter.

The questions that surround the notion of intentionality are of perennial importance for as Brentano, and later Putnam[1], have said it is a problem that will not be reduced to talk of functional states, nor will it just go away if it is confined to the realms of folk psychology. Those philosophers who side with eliminative materialism dismiss folk psychology as being unworthy to describe the natural world. With the advent of computers there was the hope that, through analogy with the functional states of a computer, the mental states of the mind, that is its intentionality, would be explained. However, the problem has remained with us.

There are two new hypotheses that arise in more contemporary work, one physicalist, that of Jerry Fodor's *Language of Thought*, and the other, the reductionist view of Daniel Dennett's *Intentional Stance.* Both of these are considered at length because they lead a way into the discussion of what criteria we generally expect for the ascription of intentional states to systems other than ourselves.

I will also look at Husserl's work because of the importance he gives to both the context and experience of the system in its attitude or stance to the world. The relation of experience to understanding will then be extended to the question of how we see ourselves in our world as individual human systems; and from this I will proceed by looking at Searle's notion that the perception of the self in relation to a particular aspect of the world is of the utmost importance for any kind of understanding. Accompanying this is the notion that it is because we can see ourselves in the world, and in so doing are reflective, that makes the human system distinct from non-human systems.

I will go on to confront one of the central issues in this area: that of identifying the conditions under which we are likely to attribute mental states to other systems. I shall examine whether or not there is a way of grading mental states so that we might say that being capable of processing information requires a lower order mental state than actually knowing what information is being processed. I will then look at the circumstance under which some people have been willing to attribute mental states to machines, and following that I will examine a hierarchical stratification that Dretske has drawn up for differentiating between the intentional states of information processing, knowing and believing. Only in the third case, of believing, does Dretske admit that the system is capable of true understanding. At this stage with the concepts of mentality, intentionality and understanding under our belt I will move on to look at the Chinese Room argument and some of the responses that have been made to it; most notably by Paul and Patricia Churchland, by Steven Harnad and by Margaret Boden.

Throughout this chapter the notion that intentionality and mental state theorising is a fairly tangled web will not fail to come across; thus the next feasible step will be to look at intentionality in relation to Aaron Sloman's work on the complexity of the internal architecture of the system to which we are making a particular attribution. In this vein I will conclude the chapter with a look at the implications that the structure of a system's architecture has on its capability to exhibit a variety of actions.

2.2. Mental states - an introduction

In the introductory chapter I was able to narrow the problem of mentality down to what have been variably described as 'mental states'. Of these states we know there to be a great number; some of which are significant only to the individual, some when considered in relation to other systems, whether organic or inorganic, and some only when interacting with another mental system. I will briefly analyse mental states as having two parts, one as experience and the other as reflexiveness. Or, more simply, the qualitative or experiential and the cognitive, such as having beliefs, knowing a fact and so on.

I shall consider here these two most prominent aspects of mental life and the question of which mental states are relevant to the abilities of the systems that have them. My overall proposal will be that the system that has the most obvious mental states, that is, that we assume to possess an active mental life, will also have the greatest capacity to act or behave. Thus the systems that I am mainly concerned with will be those that occupy the higher levels of the phylogenetic scale, and in particular those that exhibit what is usually described as 'intelligent' behaviour.

Descartes was writing in the seventeenth century about this problem and I will open up this discussion by looking at one of his most significant passages[2]; a passage that might even be read as 'a proto-refutation of the Turing test!'.[3]

Descartes was certainly keen to contrast the essential human traits with those he considered 'merely mechanical'; indeed he did wonder "if there were machines which had a likeness to our bodies and imitated our actions, inasmuch as this were morally possible"[4] would we be capable of telling them apart from 'real men'. He argues that for two reasons it would be impossible. The first is that "they could never use words or other signs, composing them as we do to declare our thoughts to others"[5] and secondly, "although they (machines) might do many things as well as, or perhaps better than, any of us, they would fail, without doubt, in others, whereby one would discover

that they did not act through knowledge, but simply through the disposition of their organs".[6] It is certainly an interesting, and perhaps even prophetic, passage.

It is possible to infer from this excerpt that the difference between the merely mechanical and human beings is that the former is not in possession of mental states. I shall now look more carefully at what definition is attributed to mental states; what they are; how they are manifest and in what way they can be attributed to other systems. This will require an examination of how propositional attitudes are used to express the way in which a linguistic entity sees itself in relation to its world and an analysis and assessment of some of the recent work in AI/philosophy that deals with epistemic states, the manipulation of symbols and the creation and attribution of semantic content to symbols.

At this stage it is useful to point out that in this section the use of the terms 'mental life' and 'mental action' will be reserved for use solely in relation to organic systems. It is only later in the major body of the thesis that I will look at whether or not it might be justifiable to extend such notions to inorganic systems.

There are a great many reasons why mental states might strike the inquirer as unusual; for a start, although there has been a great deal written about them very little of this writing tends to be in any sort of agreement. Then there is the difficulty surrounding the nature of a 'state' that has a content which cannot be isolated and specified. And finally, there is the problem with whether or not the content of the mental state is a concrete entity or something completely abstract. In this chapter I shall endeavour to straighten out some of these problems.

The first of these reasons can be easily dealt with by looking at the variety of writing that there has been and comparing them to identify instances of overlap and the areas over which there is most conflict. To begin with I will briefly look at mental states and how they relate to what are commonly describe as 'mental acts'. Then I will examine what P.T. Geach says about mental acts and their relation to propositional attitudes and intentionality. In this explication I shall take for granted that the behaviour of the organism is both mental and physical, and I will adopt the Bishop/Dennett line

that states that behaviour 'counts as action only if it is explicable in a special kind of way, namely, in terms of the agent's *reasons* for performing the behavior explained'(Bishop)[7] or that we have what are described as *intentional explanations* (Dennett).[8]

Intentional explanations of behaviour offer the reasons behind a particular behaviour, that is, '*showing the point or meaning* of what happens'[9] rather than giving a scientific explanation in terms of natural laws and probability. An intentional explanation is what is required from the agent's point of view when we inquire about the reasons behind the actions of an individual. A scientific explanation is what we get when we look into the neurophysiology or brain states of the individual.

2.3. Mental and physical acts

By using examples it is possible to make a naïve distinction between mental and physical acts. For instance, what counts as a physical action will be something like raising your arm, running a hundred metres or going to the opera, whilst a mental action can be described as a thought, such as hoping, fearing or deciding. So then, just as raising my arm is the precedent to lifting something off a shelf that is above head height, so then the mental action of deciding to make a pot of tea for my guests is the antecedent to raising my arm so that I can lift the tea-pot down from its shelf.

A mental action can also elicit another mental action. For example, if Amy has the feeling of being embarrassed on encountering someone with a moody temperament and bad behaviour, she may consequently hope that their paths do not cross very often. In this latter case her former feelings, of embarrassment, inform her subsequent mental action, that is, to hope that she is fortunate enough not to see the person frequently. Her former mental action may also suggest a coincident physical course of action like going out of her way to avoid that person in future.

Bishop argues that we have a "coapplication of intentional and natural explanations"[10], which means that events can be both agent and event caused. Our actions, both physical and mental, are problematic because from a naturalistic point of

view we want to be able to understand our actions as determined by the agent. The action needs to be described as "*agent-caused:* as determined by an agent through an exercise of that agent's control".[11] From a scientific point of view our enquiries will yield reasons that explain something in terms of being '*event-caused*'.

The sorts of explanations that we are primarily interested in here are intentional or agent-caused; and what Bishop means when he presents something as being 'coapplied' is that the action can have both an event and an agent caused explanation. So that some natural events can be brought about by the determinism of the agent whilst also having a coexistent naturalistic or scientific explanation. Bishop explains that holding this opinion affords some difficulty because the naturalistic explanation sees all events as *happenings* but from our intentional position we want to see some events as *doings*. These 'doings' are actions that have an agent, and the agent has chosen or decided to do them. So what we have are actions, of one sort or another, that are related to the mental states of the organism in a number of ways. They can be related in an entirely physical manner, as our scientific explanation would maintain; or through an ethical relation of sorts that holds the agent to be morally responsible for his or her actions.

It may seem merely tautologous to say that logically our mental states exist prior to our mental acts but in fact this also gives us some new information, namely that there must be something contained in the mental states that makes it possible for them to inform the mental acts. We might conclude from this that there are different sorts of mental action which are dependent upon the system manifesting a certain sort of mental state and from this we can infer that mental states are the precursors of mental acts and mental acts precede, perhaps even, herald mental or physical actions. It is to mental acts that I shall now look for further explication of these intricate notions.

2.3.1. Mental acts

Perhaps the clearest exposition of these issues is to be found in Geach's *Mental Acts*. In his first chapter, 'Act, Content, and Object' he deals briefly and succinctly

with our two problematic areas. He describes 'content' as the "psychological character...of mental acts". And goes on to explain that such 'content' is "expressed by the use of psychological verbs, such as 'see', 'hear', 'hope', 'think'". To make grammatical sense each of these verbs requires a noun, or "grammatical object", and so he describes them as "object-expressions". However, in anticipation of possible future problems with the word "object" Geach drops its usage and talks solely in terms of "object-expressions". He says that "such-and-such object-expressions are used in describing these mental acts; what is the logical role of these expressions?". The 'logical role' that such expressions play is to avoid making spurious references to objects or events that are believed to be actual or physical, when they in fact belong to that category of events we describe as being 'mental' or 'abstract' and which have no necessary existence in the sense of being physically 'out there'.

Although Geach is talking specifically of 'mental acts' and not, as I am doing, 'mental states' his definitions are nonetheless helpful, for a mental act will require that the organism has some particular mental states and these states must have both a content and an object of sorts. If we reverse the terms 'act' and 'state' the statement will still remain true, for being in possession of a mental state will require that the organism is acting mentally toward some 'object', whether the 'object' is 'in the mind' or 'in the world'.

The mental action being referred to is the intention to commit some action, and it is important to point out that this action can also be to ignore or store for later that incoming information which is not immediately pertinent. 'Intention' in this sense relates to the system's will to act, although the action that succeeds the intention may be a mental action and not an actual physical action. The area is now open to a discussion of 'intentionality' and how a system sees itself in relation to its world. In turn this discussion leads to an examination of the selective capability exhibited by a system when it chooses those pieces of information in its environment to which it ought to respond.

2.4. Intentionality and propositional attitudes - how they are related

So that this section can be begun with a broad idea about what is to be discussed I will refer to the four most frequently stated examples of what counts as intentionality: "(1) the fact that words, sentences, and other "representations" have *meaning*, for example, our words have meaning because we ascribe meaning to them and we then go on to use those words in consistently meaningful ways within a linguistic community that shares our understanding; (2) the fact that representations may *refer* to (be true of) some actually existing thing or each of a number of actually existing things, for example that I want someone to answer a ringing telephone; (3) the fact that representations may be *about* something which does *not* exist, for example dreaming about winning the Derby on the back of a unicorn; and (4) the fact that a state of mind may have a "state of affairs" as its object". Examples of the fourth type of intentionality would be "Ann believes that her friend is unhappy in her job" or "Arthur hopes that one day he will get a mortgage". [12]

Intentional states are described using propositional attitude statements, which contain what Geach describes as "psychological verbs".[13] Through statements of this type the individual shows itself to be in an expressible relation to the world. (It is arguable whether all relations between a subject with linguistic capabilities and an object are expressible in language, but this is not a question that I wish to enter into at present.)

Any judgement or desire we form will have to be expressed in a proposition with a predicate and what Geach calls an 'affair complex' which is representative of the relation between the subject and the object. An example would be "I believe that 'x'", where the affair-complex is my holding the belief that 'x'. It is this 'affair complex', this 'relation' or 'propositional attitude' with which I am primarily concerned here. That a system is capable of being reflexive is taken as a provisional requirement or fundamental premise of its being able to have propositional attitudes.

This idea of the affair relation between the subject and the object is by no means new, Brentano, and many others both before and after him, have talked of the importance of the relation in intentional action.[14] I shall briefly outline and address the main themes in Brentano's work and this will open up the arena for a comprehensive discussion in relation to the points identified his work.

2.4.1. Brentano's intentionality

In *Psychologie vom empirischen Standpunkt* (1874) Brentano claims that there are two sorts of phenomena, 'physical' and 'psychical'. A distinguishing feature of psychical phenomena is that they are always directed towards something. This 'directedness' is another way of describing the action of intentionality. Such acts are recognised, with reference to the affair complex above, by removal of the object-expression which renders the verb nonsensical. For instance, a wish is nothing without there being something to wish for, nor can I have just a hope with nothing as the goal of that hope.

However, an important point to note is that it is the relational activity which is a mental or psychical phenomena that is important and not the actual relation between the mind and an object, since that would entail the necessary existence of the object. Having mental directedness does not mean that the object of thought has physical existence. I can, for example, wish upon a star or wish a friend a successful and happy life.

So the focus is on the mental experience of the intentional object and such objects have what Brentano describes as 'immanent objectivity' or 'intentional inexistence'. The upshot of this is that when I wish for something there is an object whether physical or psychical that is in effect 'out there' that corresponds to my wish. Such objects have a special ontological status all of their own; that is, only they can be directed towards a goal or end that may or may not exist. Naturally the same can be said about the ontology of every object of propositional attitude statements since they express intentional relations.

Mental acts always refer to something and the mind, as a faculty of awareness, has the capacity to make judgements and have hopes, beliefs, fears and so on about these things. Thus it is a necessary feature of awareness that it always be about something. This 'aboutness' or 'immanent objectivity' can be thought of as the presence of an object to an aware subject, where the subject and the object are in an intentional relation.

Brentano distinguishes three types of these intentional relations. The first is when 'x' is present in my consciousness, that is, when I am only thinking about it. This he calls *Vorstellungen* (ideas, thoughts or mental presentations). The second relation is that of judgements about 'x', an example of which would be 'I believe that all humans are bipedal'. And the third relation is that of choosing to pursue or avoid 'x'. In this last relation an element of selectivity is present.

I find it difficult to accept these three relations as being entirely distinct. To begin with I believe that the second and third naturally rely on the first since it is not possible to choose to attend to something unless it is first present to mind. So the first relation is assumed by the other two. I also maintain that the third distinction collapses into the second because when choosing to respond in a certain way to an intentional object one is also, by definition, making a judgement about it. If I judge that a particular action is morally correct and I want to live a good life then I will most likely try to pursue that course of action. So the judgement seems to be all inclusive.

The difficulties that I have outlined against his three distinct types of intentional relation do not detract in any great way from the essential points that are being made. Firstly, Brentano has brought to mind the problems about the ontology of intentional objects, and he has emphasised the importance of the intentional relation between the subject and the intended object. Secondly, and following the philosophy of Kant, is that the mind is a faculty of awareness. And, by making the distinctions that Brentano does an important point is brought to light concerning the capability of a system with a faculty of awareness to exercise its own volition and select the objects it wishes to pursue or avoid.

So 'physical' and 'psychical' phenomena and the objects of 'psychical' phenomena have a special ontology. Thirdly there exists relational activity between the subject and its intentional object(s). Fourthly there is a directedness toward the objects of intentional behaviour which gives the objects an 'aboutness' or 'immanent objectivity'; this point could very well constitute part of the second point. Brentano's view of intentionality is an "in-the-head" relational view. For it is only through the faculty of understanding, that is itself only possible through consciousness, that an 'immanent object' is formed. As a result these objects being "in-the-head" only have an intentional existence; therefore, they are *in esse.*

The fifth point is that the mind is a faculty of awareness and I would like to extend this to say that the human mind is capable of a reflexive awareness which is unlike that which is possessed in any other system, organic or otherwise. This is going to be the line that I shall argue in my thesis.

I will now use these issues as sub-headings under which I will introduce more contemporary work that is related to the problems of intentionality, mental states, rationality, subjectivity and context dependency.

2.4.2. Brand's intentionality: The relation of the mind to its objects

In his *Mental Action Theory* Myles Brand holds a view similar to that of Brentano.[15] It is a theory concerning the relational activity of the mind to its objects and in it he states that the mental antecedent of action includes a number of mental states, 'believing' and 'wanting' to name just two. The one he says that approximates most closely to the cause of the action is 'intending'. Having the intention to act is much more determined than just wishing or hoping for the intention to act. It means that the system is now disposed to act, and it is that disposition that makes it possible to get over the hurdle that separates action from inaction.

Propositional attitudes can be thought of as the mental attitudes that are associated with the system having particular mental states. The ontology of the objects of such attitudes is ambiguous by nature as we have already seen in Brentano, but Brand

overcomes this difficulty by describing the objects of such attitudes as 'properties'. This is slightly different from the conventional sense of 'object' and 'property' when a property is something that is ascribed or belongs to a physical object. In Brand's sense it is possible to analyse any of my attitudes in terms of the propositional attitudes that take 'properties' as objects.

This is a fairly robust notion, capable of incorporating the complex relations between propositional attitudes, mental attitudes, their objects and properties; it also allows different types of attitude to be directed towards the same object, and for the same attitude to be adopted towards many different objects. These relations are very complex but it is possible to see that Brand means that one only attributes a property to something when one can first attribute to oneself the position of being in a unique relation to the state of affairs which has that property.[16]

2.4.3. Fodor's intentionality: Language of thought

The relational aspect of the affair complex is also of importance to Jerry Fodor. In chapter seven of *Representations* he argues for propositional attitudes "as relations between organisms and internal representations". He claims that his view is "probably true" because it is both "plausible a priori" and "what's demanded *ex post facto*". But, I believe he would also argue that his view is a common-sense one for it is capable of explaining a great deal more than any of the other theories that exist to date.

His view is a physicalist one which correlates the mind with the brain so that any description of intentionality can be examined by an investigation of the human cognitive faculty. The brain has a 'language of thought' in which the intentional state is encoded, which means that the cognitive function of the intentional state is literally an encoded propositional attitude statement. The brain is a 'semantic engine driven by intentional states', so that our beliefs, desires, suppositions and so on can be said to be real features of our brains.

In the head there is what counts as first order intentionality since it is the encoded propositional attitude, a feature of the brain. Using language to create and utter

propositional attitude statements is to have second order intentionality and such second order intentionality reflects or represents the actual brain states that we have. In this way I can think of events that might occur by having representations in my head of actual intentional states. All of this is possible, according to Fodor, because of our language of thought or 'mentalese' within which the propositional attitudes have their first representation. Any single propositional attitude can be applied to a variety of situations. For instance, I can say 'I believe that it is cold outside' on many different occasions, and my meaning may vary a little, but the intentional states I have are essentially the same. It is only with processing that the propositional attitude becomes shaped for a specific circumstance and no other.

It is our brain states that represent and it is these internal representations that are of greatest concern in any psychological explanation of human behaviour. The representations talked of are those of intentionality or propositional attitudes, and it follows that intentionality must be a feature of our brains that has its existence in mental states. Fodor has no time for the phenomenology of Brentano that proposed 'immanent objectivity' for the objects of propositional attitudes and what he offers instead is a computational or representational theory of mind with 'mentalese' as a descriptive language. It states that any propositional attitude is a computational relation between the system and its internal representational system. The information being represented is the object to which the propositional attitude refers. It is the information or collection of mental states that is the 'mentalese'.

It is important to note that in his theory Fodor explains that mental states are related causally because of the system's capability both to represent and to process information. However, he also states that the brain is capable of operating on both a causal and an intentional level; but to explain this I shall have to say a little about his conceptual and perceptual learning hypotheses.

Conceptual hypothesis

To understand and extrapolate from one concept to another, that is, the formation and confirmation of new concepts, the system needs to be capable of extracting the essential features of the concept and re-apply them to further instances of the concept. A significant part of this theory is that we use language to talk about our concepts. This is an idea very similar to the 'family resemblances' talked about by Wittgenstein in the *Philosophical Investigations*, except that Wittgenstein does not talk of essential features, but rather, shared commonalities.

Wittgenstein says that the only way that we can recognise something as an instance of one thing and not of another is if we can recognise the features they have in common, that is, the ways in which they are similar. This idea prompted him to think of a family group and the way that members of the family resemble one another in physical features and idiosyncratic behaviours. It is certainly the case that people say of a baby that she has her mother's eyes, or his grandfather's smile and so on; and it is in just such a way that we learn the concept 'book' or 'cat' or any number of things.

Perceptual hypothesis

Fodor says that we learn about distal objects through an interpretation that is based on a complex of proximal stimulations that we receive through our sense organs and that we build up our perceptual data through such continuing experience. It is different in a significant way from conceptual data because it is not linguistic. It might be said to be conscious, or experiential, but not self-conscious or cognitive, in the two senses of mental states that I have defined earlier.

The capacity to learn perceptually, being non-linguistic, can be shared by both human and non-human animals. But, concept learning is linguistic and for that reason it is something for which only humans have the capacity; at least according to both Fodor and Kant. The shared nature of perceptual learning, through the common feature that the human and non-human animals share of being representational systems, ought to emphasise that the language of thought is not simply an internal natural language. In

general the brain is perceptual and the difference between human and non-human systems is that non-human systems are only capable of perceptual learning, whilst human systems are capable of perceptual learning and also of the language of thought.

According to Fodor such languages have to be, in some sense, innate for "One cannot learn a language unless one has a language". Here Fodor is following in the rationalist tradition often associated with Chomsky for Chomsky claims that different languages use the same formal operations, 'universal grammar', for the generation of 'well-formed' sentences. To make this possible he states that all children have to be 'endowed with an innate capacity' to use the universal grammar that makes it possible for them to learn the language of their environment. Fodor says essentially the same thing when he says that to learn a language requires that we have a prior capacity to grasp the formal operations needed in order to use a natural language. The difference between his view and Chomsky's is that Fodor stipulates that we have at least two languages already 'wired in': namely the language of thought and the perceptual language that allows us to interpret raw sensory information.

Fodor makes use of a computer analogy

The machine's internal language is a private language, but a programming language is a public language for it is the language with which the programmer communicates with the computer. For the 'innate' component Fodor offers the machine language compiler which gives the computer the capability to interpret the rules and functions of the programming language. Fodor then goes on to equate the compiler with the human representational system that is present in each potential language user. But this argument seems all too easy and I find myself puzzled about such an analogy that freely compares the human cognitive capacity, that we know all too little about, with the computer's capability to follow rules that it has been given and which cannot after all be 'innate'. Something fundamental seems to be missing and, I would argue, it is the element of understanding that is talked of everywhere from Frege to Searle. Human systems are capable of understanding the reasons for their actions and when acting

consciously they are capable of 'grasping' meanings and applying them abstractly, something I think Fodor's computer would find impossible to do.

That Fodor's attitudes have been equated with the theories behind 'folk psychology' is due to his maintenance that the actions of the individual can be explained by reference to his or her beliefs. From this claim it is reasonable to expect that Fodor would also believe that all behaviour can be explained in terms of the totality of the individual's propositional attitudes. But once again, there is something missing for nowhere does Fodor talk of the consciousness of the organic system. Perhaps then Fodor would wish to conclude that the totality of propositional attitudes would be enough to explain our conscious behaviour, and if this is so would Fodor also wish to accept that the totality of propositional attitudes can also explain our self-conscious behaviour? I suggest that he would not for this is a very tall order and not one that he can hope to fulfil by simply examining the individual's 'language of thought'. Indeed he would encounter a new set of difficulties when he would come to explain the sort of animal consciousness that gives all the indications of being reflexive, for which there can be no recourse to a 'language of thought'.

2.4.4. Dennett's intentionality: The intentional stance

In this same area, but in contrast with the work of Fodor and Chomsky, Dennett has proposed the adoption of the *Intentional Stance*. This is certainly one of the most interesting theories to be proposed in recent years. Very broadly the claim he is making is similar to Brentano's in the sense that intentional states are relational but they are not 'in-the-head' relational in the way that Fodor would argue.

Dennett takes this view a step further and adds that by adopting an instrumentalist approach to intentionality, which claims that the behaviour of a system can be explained, predicted and controlled solely by the ascription to it of beliefs, goals and rationality, one can also ascribe intentionality to systems that are not organic. When taken to its logical conclusion this stance permits us to describe some already existing computer programs as intentional systems; for anything that can have its behaviour

predicted by attributing to it both propositional attitudes and rationality, is, per se, an 'intentional system'.

When manifesting a mental state, that can be described using a specific propositional attitude, the sentence or statement of the propositional attitude is not somewhere embedded or represented in processes that are in our heads. Beliefs are mental attributions that we apply to the propositional attitudes that we use as descriptions of states of affairs that we encounter in our interaction with our worlds. They are, very simply, abstract notions that we use for predicting the behaviour of other organisms and systems that surround us. Dennett's view is a non-reductionist account that does not require that propositional attitude notions be reducible to anything that can be stated in physicalist or functionalist language. Because of this he fails to look at the nature, ontology and causal powers of propositional attitudes. Nevertheless it is still a valuable basis from which to begin an examination of propositional attitudes and intentionality.

Like Fodor, Dennett makes use of a computer analogy

In *Brainstorms* Dennett offers a view of extreme functionalism where he states that the mind is to the brain as the software of the computer is to its hardware; and so that we are in no doubt about his position in this 'battle' he says on page one of chapter one, *The Intentional Stance*, "the brain (which, after all, *is* the mind)". But the computer analogy is not one that Dennett welcomes with open arms for he goes on to argue that it is really most unlikely that every human being will share an identical "evolutionally-produced program". Clearly, Dennett believes that an objective account of both intentionality and consciousness is possible for he asserts a desire for a demystification of such notions. This is a view that directly opposes that of Thomas Nagel, who states that 'the particular point of view, or type of point of view' is an absolute necessity if we are seeking a full account of reality.

It is certainly practical to admit the internal functional states of humans but it is not in our best interests, at least according to Dennett, to imagine that a one-to-one

correspondence exists between the described state and the mental process or brain state. By implication it is easy to link the actions of the individual with his or her brain states, but it is not yet possible for us to have any direct empirical evidence of the actual brain state at the moment of being in a mental state of having for example, a belief, hope or desire.

The convenience of propositional attitude attribution

The sort of functionalism proposed by Dennett allows for the attribution of the same belief state to more than one person because the attribution is not done on a neurophysiological basis, but rather on a basis of the observation of behaviour in relation to a set of events in the world. So from the observation of perceptual input and behavioural output it is possible to describe a person as being in a certain state or states of mind. In this way it is merely a descriptive convenience for us to attribute mental functions like propositional attitudes.

Viewing objects external to us, both organic and inorganic, as having propositional attitudes is convenient since it is just an extension of how we deal with our own interaction with the world. When I examine my own relation with my world it is through my mental model of my environment and my interaction therein; I build plans for the future by relating this model to my beliefs, hopes and desires and combining this view with a rational approach to what is realistically possible.

The attribution of propositional attitudes is done by observing the perceptual input of the system in a particular environment and combining this with the mental states we believe it to have. By then associating this with the assumed rationality of the system it should be possible to predict its behaviour. Such a stance can be adopted towards non-human animals, and even towards inorganic systems, and still be seen to work. For instance, it is possible to anticipate the future behaviour of an animal by watching it interact with its environment and relating this to its previous action in similar circumstances.

A disposition to rational behaviour

This is a theory that Dennett describes as "holistic logical behaviourism".[17] All the intentional language we use is replete with information about the system, its perceived relation to the world and the predicted behaviour of that system. It is true to say that when an identical piece of information is received by different people it is received and processed in many distinct ways so that each interpretation is going to be unique. The commonality between each person with that belief is that they will exhibit predictable and rational behavioural dispositions. So what we are, in effect, doing is classifying systems in accordance with their exhibited disposition for rational behaviour. This, in turn, allows us to conform to objective regularities that can be described using extensional language whilst avoiding the snags and pitfalls of an intentional language.

In *Brainstorms* [18] Dennett tells us that by adopting the intentional stance towards the objects in our world we are taking, at the very least, the "pragmatic" option. For, as he so often reminds us, it is only through such a stance that we can continue to make reliable judgements about the prospective action of the things with which we interact. The justification for this theory seems simply to be that it happens to work. If we choose not to adopt the stance we will be in a continual state of flux because so much of our action depends upon the action we think others will take. We would no longer be able to plan our actions in accordance with that of other organisms. In the words of Thomas Hobbes our lives would be "solitary, poore, nasty, brutish and short"![19]

To adopt the 'intentional stance', then, is to accept a strategy for attributing propositional attitudes to a system and predicting that system's behaviour depending on what it would be rational for that agent to do given his or her propositional attitudes. The system can be organic or inorganic, and as long as its future performance can be predicted, and thus explained, it counts as an intentional system.

The influence of folk psychology and folk physics

There are two areas to which we look for an account of our world, namely: folk psychology and folk physics. Although Dennett would want to argue that our mental

states do not have a determinate content, (contrary to both Searle and Dretske), that is that we do not possess "intrinsic intentionality"[20] or a determinate thought-content, he acquiesces in the view that elements of both accounts may be innate. However, he still maintains that for the most part they will be learned through experience. Having found that there are areas of folk physics that are counter-intuitive it can hardly be beyond our comprehension that some areas of folk psychology might be vulnerable to further empirical research.

The attribution of belief can be objective or subjective. The latter, an interpretationist account, is open to cultural influence and therefore more problematic than the former, realist account. In Fodor we can see an example of the realist point of view for he states that beliefs are objective things in the head and in principle such states can be identified by physiological psychology. The interpretationist account views the attribution of belief states as being controversial in the same way that one would think it contentious to assert that some individual was deceitful.

Dennett attempts to meld both positions by claiming that although belief is an objective phenomenon it can be better understood by adopting the interpretationist's predictive strategy, the *intentional stance*. Anything that can be said to have beliefs, and be described as a true believer, is, in Dennett's opinion, an intentional system. To adopt the intentional stance one must first treat the system whose behaviour is to be predicted as a rational agent. Given that the system is in the world and that it will want to further its goals, by adding the attribute of rationality it should be possible to predict its actions. If we simply work from a folk psychologist premise it is possible to extend the notion of rationality to other systems if we observe enough of their input and output states and compare their interactions in the world with our own rational behaviour in similar circumstances. By adopting the intentional stance it is possible to attribute propositional attitudes to systems other than ourselves on the very same basis.

Propositional attitude psychology is troublesome

Because Dennett attempts to draw objective and subjective attribution together we can conclude two things: 1) there is no unified, reliable view of propositions and propositional attitudes, and 2) language-of-thought psychology yields no worthwhile results. I would like now to say a little more about both these in the light of Dennett's proposed *notional attitude psychology.*

In chapter 5 of *The Intentional Stance*, 'Beyond Belief', Dennett tries to do away with the whole troublesome area of propositional attitude psychology. Propositional attitudes can be analysed into three variable components; "X [subject] believes [attitude] that *p* [proposition]". When enquiring into the nature of the proposition we find that three quite distinct views are held. The first says that propositions are like sentences, that is, symbols that are held together in a syntactical form. The second view claims that propositions are just sets of possible worlds and the third states that propositions are ordered sets of objects and properties in the world.

That three views exist is a mark of the number and complexity of conditions that they are required to meet. Propositions have to be bearers of truth-value, so that we can say of something that it is a true or false statement. Next they have to fit the requirements of an intensional language, that is, that they have to be able to cope with referential opacity;[21] and finally they have to have a 'graspable' meaning.

Dennett argues that in the light of the work of people such as Kaplan, Perry, Putnam and Stich, it is not possible to fulfil all three of these conditions at any one time. In the face of such opposition the only retreat would seem to be into *sentential attitude psychology*, which is a language-of-thought hypothesis. Dennett describes four approaches that lead to just such an hypothesis.

The first is that sentences 'in the head' are in some sense physically "grasped" when we think of an abstract proposition. Secondly, sentences about objects must be composed of symbols that represent these objects since they cannot be composed of the objects themselves. The third approach is that whatever the sentences are in our heads

they must be able to account for the problems posed by referential opacity. Lastly, of sentences having content and syntax, sentences in the head are supposed to have syntax.

By following these four approaches the hope is that more can be learnt about whatever propositional attitudes are held by the system. However, the theory runs into a number of problems. To begin with, and as mentioned above, it is more than doubtful that any two people could ever have the same language-of-thought[22], and it is therefore very unlikely that any two people could have precisely the same beliefs. With this in mind it is clear that sentential attitude psychology is trying to distinguish too precisely between different psychological states. The next problem is that it is already presupposing that it is possible to access the syntax of propositional attitudes before being able to know their 'semantic' properties. A final criticism is that it assumes that it is possible to put semantics into a verbal form and this may turn out not to be the case even when we have more information available.

Between the 'language of thought' and the environment

Dennett puts forward a 'coping' strategy that is intermediate between the language-of-thought and the external environment of the organism. It is called *notional attitude psychology* and it is not constrained by any hypothesis about internal representations or where such representations (if they were to exist) would be located. Notional attitudes are the constituents of the system's "notional world"; and the notional world is the world at that time and that place that the organism is best equipped to deal with. This theory offers one noticeable advantage; namely, were I substituted for a person identical to me in a world identical to my present one, then I would possess all the relevant belief states without having had any interaction in the "Twin Earth". This eases a small proportion of the problems associated with possible worlds. The other advantage this theory has is that it does bring to light the difficulties associated with adopting the reductionist language-of-thought hypothesis.

A similar notion can be seen in the "bracketing off" feature of Husserl's work in phenomenology and the phenomenological reduction (1931) - see next section, 2.4.5., also in Quine's theory of the indeterminacy of radical translation (1960) and in more recent work by Searle on "Aspectual Shape" (1990).

2.4.5. Husserl's intentionality: Our experience is what matters

Initially, Husserl would say that we are aware that the world is 'spread out in space, endlessly becoming and having endlessly become in time'.[23] Simply said, the world and everything we perceive is out there whether we choose to attend to it or not. We have, what Husserl describes as, a "*natural attitude*" which enables us to observe our world, have feelings with regard to our world, to make judgements about our world and to resolve to act in relation to our world. "Moreover, this world is there for me not only as a world of mere things, but also with the same immediacy as a *world of objects with values, a world of goods, a practical world*".[24]

Spatially most of my world remains within an area of indeterminacy, a bit like my peripheral vision; "my indeterminate surroundings are infinite, the misty and never fully determinable horizon is necessarily there". So too with my temporal perception; "this world, has its two-sidedly infinite temporal horizon, its known and unknown, immediately living and lifeless past and future".[25] My natural attitude in the world permits me to "change my standpoint in space and time, turn my regard in this or that direction, forwards or backwards in time; I can always obtain new perceptions and presentiations, more or less clear and more or less rich in content, or else more or less clear images in which I illustrate to myself intuitionally what is possible or likely within the fixed forms of a spatial and temporal world".[26]

This section alone portrays the whole richness of the human mental ability, for in it we can recognise many of our higher cognitive abilities. The individual has a continuous array of perceptual inputs that are its source of new information, from which it can select the most important things for immediate attention, and with

consideration of past events through a richness of mental representations, decide what future action would be to its advantage.

The world is continually present for me, even when I focus on some abstract concept like mathematics or logic. My standpoint to the world is then a logical or mathematical one and the background to my consciousness of mathematics is my natural attitude to the world. I am said to be in an 'arithmetical' or 'logical' attitude. Being capable of a phenomenological reduction means that a complex system can bracket off sections of its world in favour of emphasising other more abstract interpretational stances.

I think a word or two ought to be said here about 'bracketing' for it is a complex term used by Husserl in the philosophical context of the 'phenomenological reduction' and not one that is immediately clear. The best analogy I can think of is with parentheses. If I read a sentence from a paragraph and within that sentence there is a set of parentheses, I will first read the sentence for its meaning by ignoring the information that is in the parentheses. When I feel that I have a complete understanding of the sentence I will go back and read the sentence again this time incorporating the information that is inside the brackets. On the first reading I will have 'bracketed off' information that is not immediately necessary for an understanding of the sentence. By the time I have made a second reading I will have understood the information contained in the first sentence and also the non-essential information that was enclosed in the brackets.

This example corresponds well to Husserl's meaning of 'bracketing', for what we have done in our sentence is select the information that is of immediate interest to us and ignore the non-essential or superfluous information stored in the brackets; and what Husserl suggests we do in our perceptual environment is select the information that is immediately relevant to our situation, attend to it, and 'bracket off' or ignore the information from our environment that is unnecessary to us at that moment. So when working with logic or mathematics my experience is structured with the abstract world coming to the fore and the natural world receding into the background. Both worlds are

related to the ego or consciousness whilst still somehow being distinct entities. There is an element of choice now evident in choosing what I will or will not attend to from my perceptual information.

This type of experiential structure is essentially the same for everyone. But the element that will change is the content of our experience that will vary from person to person. The one bit that is common to everyone is the objective spatio-temporal world to which we belong, that is, our 'natural standpoint' of phenomena or physical things. The area of personal experience is the individual's private perceptions.

By *bracketing* off the belief we have in the totality of objects and events and instead concentrating on the private, inner or 'noumenal' experience we have of them we are performing what Husserl describes as the *'phenomenological reduction'*. We literally reduce our world of phenomena until we reach our subjective experience of particular phenomena. Things in the world still exist but we consciously refrain from making judgements about them. In the sense explained above, we 'parenthesize' phenomena in the world and look instead at our experience of the relation between us and the world.

The next step is to try and describe this process of experiencing and look at what sorts of structures are left outside the 'brackets'. Structures of this sort, our experiences, are called the 'forms of consciousness' and it is only through them that mental experiences are possible. A most important point is that these experiences are not just of other objects and states of affairs, but also of the personal *'transcendent ego'*; the so-called *'Archimedean point'*. Such pure consciousness is arrived at after bracketing when the phenomenological reduction is complete.

Intentionality as a principal theme of phenomenology

Intentionality is the general theme of 'objectively ', or object-oriented phenomenology. It is shared by all systems with mentality since intentionality is a characteristic of consciousness. 'Intentionality', then, is the term that Husserl favours for describing the experiential or *'phenomenological'* structures that are left after we have successfully bracketed off the way we naturally view our world in space and time.

However, Husserl admits that the manifestation of intentionality will alter with regard to the differences in mental structure of different types of system; 'we cannot say of *each* mental process that it has intentionality in the same sense'. So that the mental structure of a dog or cat will influence the character of the intentionality they possess, if indeed they possess any at all. In intentional behaviour we are 'conscious *of* something' and regardless of the existence or non-existence of the 'object' there will be some correlation between it and our intentional behaviour towards it.

The importance of context in phenomenology

Our intentional attitude to something, i.e. our hoping that *x* or believing that *y*, lacks objectivity since it is in our experiencing and not the experience of the physical thing. It can only constitute 'intuitive appearances of objects' and not the objects themselves. It is our attitude that is important because it dictates the context within which our experience takes place and is examined; therefore it is our attitude which is intentional.

I would like to place emphasis on a couple of things of importance in Husserl's work. The first is the stress that he places on his work being conceptual, or as he describes it himself, "eidetic". Husserl's concern is with the notions or acts of believing, hoping and perceiving when all else is removed and not with the intentional objects that we each associate privately with the notions.

A second aspect which is of significance is that of the 'conceptual attitude' that I have towards something, so for instance, when I am conscious of some logical thing I will adopt a logical stance for my understanding. Just as with Dennett's theory when I am trying to describe and predict the behaviour of another person I adopt the 'intentional stance', so for example, for Husserl when I am thinking of some ethical matter I adopt an 'ethical stance'. I bracket off my natural standpoint and think in 'ethical' terms about the problem, thus enabling me to give the matter my whole concentration. What becomes of importance is the context in which I examine my experience. It has been proposed that 'languages-of-thought' other than the two already

discussed by Fodor[27], might exist, and so too there might be an infinite number of 'experiential stances' open to a complex system; but to have an understanding of any of them we need to look carefully at the experiential context.

2.5. Searle's intentionality: Experiential context

In this section I will look at the importance of contextual aspect in a paper by Searle in which he examines the notion of 'aspectual shape'. Just as Husserl talks about the bracketing off of the natural stance and the adoption of a specific conceptual standpoint in dealing with a particular area of enquiry, so too Searle talks of the importance of our perceptual, or even personal context in the way we look at the world. Very simply, the idea behind 'aspectual shape' is that our perception or thinking is always from one particular point of view or aspect whether spatial or contextual. As Searle says "Whenever we perceive anything or think about anything, it is always under some aspects and not others that we perceive or think about anything."[28]

Aspectual features are those that are perceived under a particular aspect, and it is they that make an intentional state into a mental state. The aspectual feature can be seen as a relation of some type between my experience and the neurophysiology that makes up my brain states. Under aspectual shape my experience of a butterfly is a conscious experience that I have under a specific point of view. My experience of the butterfly has some features which are essential to it and to it alone. "Every belief and every desire, and indeed every intentional phenomenon, has an aspectual shape."[29] So what we have with aspectual shape is the conscious experience of a thought, an object or a state of affairs.

In Husserl's phenomenology the two important features are, first, that there is a relational aspect through which we examine our experience of our belief or some other form of intentionality; and second, the conceptual attitude under which we interpret our incoming information. 'Aspectual shape' encompasses both of these for Searle. Under it we have experience of the world and the experience is had from our own particular point of view. For instance, my experience is had under an aspect that is specific to me.

Aspectual shape is, according to Searle, an essential feature of all our intentional states; in such a way that intentional states only become conscious mental states because they possess these aspectual features. The aspectual shape is important because "it constitutes the way the agent thinks about or experiences a subject matter".[30] So aspectual shape is the thinking or experiencing of the object of our intentions.

To overcome any spurious ascription of intentionality Searle proposes a bipartite distinction between intrinsic and 'as-if' intentionality. The former, he claims, is that which is applied to those things which we know possess mental states, whilst the latter is a form of metaphor attribution applied to those things which have no mentality. An example of as-if intentionality is "The thermostat on the wall *perceives* changes in the temperature".[31] The thermostat does not actually perceive but its action makes it look to us 'as-if' it perceives.

Searle sees Dennett's intentional stance as a denial of 'as-if' intentionality because Dennett proposes that we adopt the intentional stance with objects that do not possess mental states. According to Searle if we accept Dennett's thesis then we must also accept that everything in the universe is mental. Under this reading of Dennett the adoption of the intentional stance entails that we also accept a position of true panpsychism!

Of 'deep unconscious mental intentional phenomena"

A problem arises for Searle, and it is this: unconscious intentional states exist only as a matter for "third person, objective, neurophysiological phenomena" even though all our intentional states are supposed to have aspectual shape and aspectual shape is not meant to exist at the level of neurons and synapses. Searle frees himself from this quandary by proposing that all unconscious intentional states are "*potentially* conscious", " they are possible contents of consciousness".[32] "When we describe something as an unconscious intentional state we are characterizing an objective *ontology* in virtue of its *causal* capacity to produce consciousness."[33] What we are left

with is the fact that "any unconscious intentional state is the sort of thing that is in principle accessible to consciousness".[34]

Searle argues that there are no unconscious intentional phenomena that are not, at least potentially, conscious. All intentional states have to possess aspectual shape if they are to be mental states at all. So where there is no aspectual shape there can be no intentional phenomena; "but where there is no fact of the matter about aspectual shape there is no aspectual shape, and where there is no aspectual shape there is no intentionality".[35] The conclusion then is that unconscious states, such as Chomsky's 'innate grammar', which are not 'in principle accessible to consciousness and are what Searle describes as "deep unconscious mental intentional phenomena"[36], do not exist.

What are we left with as notions of unconscious mental states? Firstly we have the "as-if metaphorical attributions of intentionality to the brain which are not to be taken literally"; then we have "shallow unconscious desires, beliefs, and so forth", like "repressed consciousness"; and thirdly there are "shallow unconscious mental phenomena which just do not happen to form the content of my consciousness at any given point in time".[37] All of these sorts of unconscious phenomena are, at least, potential states of consciousness.

What evidence is there for intentional states?

For Searle behaviour of a system is not sufficient to demonstrate the relation between its neurophysiology and its intentional states. "Behavioral evidence concerning the existence of mental states, including even evidence concerning the causation of a person's behavior, no matter how complete, always leaves the aspectual character of intentional states underdetermined."[38] It is always possible to infer from epistemic grounds, in this case neuronal firing, that something is present and from that presence infer the concomitant existence of something. But, such inference alone does not suggest a strong justification for the ontology of a particular mental state. The behaviour without reference to the relational aspect of consciousness is not enough. It is not possible to determine that what someone else means by "water" is the same as

what I mean by "water". They may mean a particular chemical compound of hydrogen dioxide and I might mean the liquid I drink to quench my thirst.[39]

There can be no "lawlike connection that would enable us to infer from our observations" of the person and their language that what we are both referring to is the same thing. Nor is their any "lawlike connection that would enable us to infer from our observations of the neural architecture and neuron firings that they were"[40] the conscious realizations of a particular desire for something or a certain wish that something else.

Thus Searle concludes that; (i) all intentional states must have aspectual shape, (ii) that all unconscious mental states are in principle accessible to consciousness, and (iii) that it is not possible to infer from the knowledge of one thing, in this case neurophysiological states, the ontology of another thing, and again in this case that a particular mental state exists.

2.6. The ability to understand and how we see ourselves in the world

To summarise our findings so far; according to Husserl we adopt an interpretational stance that depends upon our immediate surroundings and our ability to withhold the natural standpoint with which we interpret our interactions with our everyday physical world. The world within which we adopt a particular interpretation is that world within which we have exercised the phenomenological reduction and extracted our sense data in favour of a 'return to the facts' of simple experiencing.

In Searle we have the notion of 'aspectual shape' which is necessary for every intentional state to become a mental state. It is under this aspectual shape that we have thoughts or perceptions under a specific aspect or interpretation. This is just another way of saying that our perceptual relation to a particular state of affairs is the one that is important for our understanding of our context *in toto*.

In Husserl we have a contextual standpoint when our natural world is 'bracketed-off' and Dennett argues that the system's external state of affairs is important because the intentional stance relation is not 'in-the-head', but, is instead, related to what we

interact with in our environment. Searle considers context to be important for our mental representations to be able to represent; and anything with intrinsic intentionality can understand its context in a way specific to it.

For Wittgenstein, too, context is most important for understanding language. By its very nature our language has a public sense since it is used in a public forum. My own language can have a private reference but it is of no use for conversing with other human beings unless we share a commonality of senses for our words. Words like 'pain' and 'sadness' have a private reference, for only I can know what it feels like when I feel sad or I have a toothache; however, such words have a shared public sense from which we can understand the life of another human being.

The context is all important for an understanding of this type of 'private reference' word. For example, if a little girl falls, grazes her knee and begins to cry saying, at the same time, that she is hurting, then all the circumstantial evidence would suggest to the observer, who cannot share in her pain, that she is indeed in pain and that her pain behaviour is understandable from her fall. However, of the little girl who says that she is too sick to go to school, and we know that she has a spelling test that day, the evidence, or context, would suggest that her behaviour is in fact a ruse to avoid the unpleasant test.

Colour words are also subjective, or 'private', but again we share their meanings in a common context. For if I use the word 'red' to describe a London bus, you can say that you know what I mean by 'red'. If I start talking of my favourite shades of colour then you might have difficulty understanding exactly what it is I mean. But from both examples it can be seen that it is possible to talk of, and understand, another persons intentional language by observing their behaviour and in so doing sharing in the context within which they are using their language.

Hubert Dreyfus also considers context to be of great importance. In fact he argues that context is essential for cognition to occur at all and that contextual considerations, along with the appropriate social behaviour, are required for a complete understanding of the cognition of the organism. It is, for Dreyfus, more a matter of knowing what to

do in a particular social context than knowing what proposition is the most appropriate given the situation. Our skills, as mental systems, are acquired and learnt through repetition within a specific social context and it would, for Dreyfus, seem incredible that a non-mental system could acquire the appropriate social skills in just the same way. In much the same way as Searle's aspectual shape is seen to exist as an important part of our experience Dreyfus claims that in our personal cognition of our world an essential role is played by our embodiment of our perception in our interaction with the world. What we can see emphasised here is the importance of the experience being 'ours', i.e. that we can see ourselves in relation to our experience of our world.

So, just as Wittgenstein and Husserl did before them, Dreyfus and Searle argue that social and cultural surroundings are of the utmost importance to our understanding of the world; "all intelligent behaviour must be traced back to our sense of what we *are*," and we are social animals that have linguistic and non-linguistic interrelations with other social animals in our world. What is now of importance is how we move from this state of understanding to intentional states, like knowledge and belief.

2.7. The attribution of intentional states

'Understanding' and 'knowing' are important aspects of mentality that will recur throughout this thesis and I would like, for the moment, to examine the relation that Wittgenstein sets up between them. For Wittgenstein having, or claiming to have, an understanding of a state of affairs is very closely related to saying 'I can'. "The characteristic of words like "understand" and "can" is that they are used alternately for (a) something occurring in the mind as a conscious event, (b) a disposition, and (c) a translation."[41] Of most interest to us now is (b) for it is the use in this sense that "overlaps with 'is able to'"[42], for this is the sense in which understanding is linked to a disposition to perform a particular act or set of actions. An accurate interpretation of information is "illustrated by one's being able to do a certain thing when one understands".[43]

For instance, in the *Philosophical Investigations* Wittgenstein explains that knowing something is not the same as having certain epistemic states but rather it is a demonstration of an ability to 'go on'. We use "Now I know!", "Now I understand!" and "Now I can do it!" interchangeably to mean "Now I can go on!".[44] Again in paragraph 151 in the example of the sequence '1, 5, 11, 19, 29' that A was writing, if B knows how to go on he will respond with a statement like "Now I know how to go on". What B has admitted is that if required he believes he is capable of giving behavioural evidence for his understanding of the sequence.

So understanding in this sense is actually that you know the answer and would be able to offer an objective account based on previous learning from other like experiences. It is only through an exhibition of the correct understanding behaviour, that is, that B answers '41', that we can say of him that he has understood in this context. When he does respond in the correct way we can say of him that he does 'know'. It is not possible to say of someone that they know that something is the case when they have claimed to understand but failed to show that they know by actually 'going on' and offering proof. So the criteria for attribution of epistemic states is not just that the system offers behavioural evidence but that the behavioural evidence it offers is *correct*. For instance, if B says that he has understood the sequence but then he proceeds in the sequence by saying '42' then he quite obviously has not understood and it would be wrong to attribute a knowledge state to him.

If we look closely at these criteria they seem to be the same as the ones we use for the attribution of mental states, whether rightly or wrongly, to other non-human beings. The distinction Searle makes between intrinsic intentionality and 'as-if' intentionality bears this out, for we say of the lawn that it is 'thirsty' or of the thermostat that it 'perceives', it is simply the attribution of intentional states based on what *seems* to be appropriate and correct behaviour in the system's context.

The attribution of 'as-if' intentionality is the result of our treating other non-human systems as though they had a mental life similar to that possessed by human-beings. It is a sort of silicon or mechanical anthropomorphism that allows me, without too many

raised eyebrows, to say of a non-mental object like my computer that it 'runs quickly' or 'plays up' when it knows I am in a hurry. It is an over-extension of metaphors or use of simile that we understand, from Searle, to be used in connection with non-mental systems. He would argue that once we go beyond realising this specific use and exaggerating the capabilities of these systems, in other words attributing to them a mental life that they do not possess, our use is erroneous and we are misleading ourselves.

2.7.1. Intentional states attributed to machines

McCarthy's views

McCarthy gets into deep water in just such a way; for instance, he says "To ascribe certain 'beliefs', 'knowledge', 'free will', 'intentions', 'consciousness', 'abilities' or 'wants' to a machine or computer program is legitimate when such an ascription expresses the same information about the machine that it expresses about a person".[45] So it seems that he is satisfied, indeed even blasé, about the ascription of mental qualities to non-mental systems.

McCarthy argues that it is "useful" to ascribe intentional states to machines because the ascription may be able to help us "understand the structure of the machine, its past or future behaviour, or how to repair or improve it".[46] The same ascription may also give us information about the "limitations on our own ability to acquire knowledge".[47] With this in mind we should have the accumulated advantage of being able to predict the future states of the machine on the basis of what we know about its previous states and present structure.

This is an argument that shares many similarities with Dennett's reasons for adopting an intentional stance towards objects, both mental and non-mental, in our world. Indeed, McCarthy does say that he "emphasizes criteria for ascribing particular mental qualities to particular machines rather than the general proposition that mental qualities may be ascribed",[48] which is what Dennett is doing in *The Intentional Stance*.

McCarthy also wants to argue that it is easiest to attribute these mental qualities to simple machines of which we know the structure; as he says, "machines as simple as thermostats can be said to have beliefs, and having beliefs seems to be characteristic of most machines capable of problem solving performance".[49] What this suggests is that, if we observe a system in one state and it then performs an action that places it in another state, then we should attribute to it all the analogous intentional states that we associate with the system that has a full and active mental life.

But I think McCarthy goes too far, for it is as though he really does consider machines to possess mental lives, however limited those lives may be. Examples of such thinking are prevalent, an obvious example is when he says that "present machines have rather varied little minds"[50] and he then ventures to state that our human capacities to love and hate are programmable, and although slightly more difficult "mental qualities like humour and appreciation of beauty" are not beyond being modelled.

Rosenschein's views: The machine innards considered

Rosenschein takes up a similar position to McCarthy's except that Rosenschein looks at the physical construction of the machine, and is concerned particularly with the ascription of knowledge states. One further difference is that Rosenschein first admits that the concept of knowledge as talked about in AI rests "on a very limited conception of what it means for a machine to know a proposition".[51] For him life is certainly more straightforward for he only accepts that a machine knows something if the state can be encoded in the form of a formal language sentence, or that the sentence can be derived using the "rules of an appropriate logical system".[52]

The human approach

In a technical note of 1985 Rosenschein contrasts the 'interpreted-symbolic-structure' (ISS) approach with that of the 'situated-automata' (SA). In the first case the state of the machine is that of an encoder of symbolic items of data encoded by the interpreter. The symbols are so called because they map pieces of the internal state of

the machine onto pieces of world state. Using this approach it is possible to ask "What information about the world is encoded in the state of the machine?"; or in the language of intentionality "What does the machine know?".

For the ascription of knowledge this approach requires "viewing the machine's state as structured in a certain way; knowledge is not an objective property of the way the machine is embedded in the world".[53] In the ISS approach knowledge ascription depends very much on the wishes of the designer. In fact so much so that if she wishes to assign a different interpretation to the same symbols, (its overall structure), the machine will be said to be in possession of different knowledge. Nothing else in the machine or in the world needs to change for this to be the case.

The mechanical approach

However, Rosenschein is not satisfied with the ISS approach because he believes that the environment and the machine's location in the environment is very important and the ISS approach does not take this into account. If the machine is to have epistemic properties assigned then it must be able to have an internal representation of an external event. So the SA approach is devised in order that knowledge can be analysed in terms of relations between the machine state and the state of its environment. This suggests that the notion of knowledge is grounded in an objective correlation(s) between machine states and world states.

There is an initial assumption that the machine is part of an environment that can be in any one of a large number of variable states The environment generates the inputs for the machine and responds to its outputs. However, the machine can only know of the environment through its direct inputs and it is feasible that some information will not reach it because is not capable of detecting it or it is unable to discriminate cases when *p* holds from cases when it does not.

In the traditional AI approach the machine manipulates the data structures that encode language in a series of logical assertions. In the SA approach "logical assertions are not part of the machine's knowledge base, nor are they formally manipulated by the

machine in any way".[54] The assertions "are framed in the metalanguage of the designer",[55] and they are used to express the underlying assumptions made by the designer or programmer and offer characterisations of information content of the machine states being designed. The programmer has to be able to "comprehend the emerging design and verify that the machine will behave as desired".[56]

Many computer applications involve the system in a continually changing physical environment and the primary task of the computer is to monitor and respond to the alterations in these environmental conditions. For the computer to do this successfully it has to be able to recognise the appropriate stimuli in its environment and from this make an estimate of the responses that would be most probable. In this case, if the machine is to be said to know that such and such is the case, then the state of the machine must be capable of mapping the state in the environment and using the information appropriately.

Thus in his proposal of a "correlational definition of knowledge" Rosenschein goes deep into the innards of the machine so that when they, the innards, are in a particular state, say state *A*, it will be occupying a specific epistemic state, but if we then adjust the connections between the machine's wires, nodes and so on. it will occupy a different epistemic state, this time state *A'*. So there is a direct relation between the internal mechanism of the computer and the state the computer can be said to be in; and Rosenschein would like to call the properties of these states, that is, the internal structural organisation, 'epistemic'.[57] It follows that the machine can be said to know something in, at least, a primitive sense, if it reflects a real world state. For a barometer to indicate a change in pressure it can be said to know of the change if and only if the true state in the world is one of pressure change. If it indicates a change of pressure when there has not been one then it cannot be said to know anything about the world. Similarly in Wittgenstein's example of being able to 'go on', if the person 'goes on' with the wrong answer they cannot be said 'to know' after all.

So in a somewhat different sense Rosenschein has admitted the importance of context for the attribution of intentional states, and like McCarthy, he ascribes mental

states to non-mental systems, and like Dennett and McCarthy he emphasises the predictive capability that the designer has when she knows the information content of the machine state, that is, what the machine can be said in a primitive sense to know. All three of them argue that once we can recognise analogous states between mental and non-mental systems, and we attribute the same intentional states to them both, it will be easier to predict the forthcoming actions of that system.

2.8. A new, computational, theory of intentionality: Dretske

Dretske adopts a novel approach to the problem of ascribing intentional states in what is described as an 'information-theoretic' account that combines aspects of the McCarthy and Rosenschein theories. In the account he ascribes mental states to a variety of systems that range from the non-mental to the mental. Dretske starts by using the concept of "*belief* to distinguish genuine cognitive systems from mere processors of information".[58] He gives the example of an information processor as something like a tape recorder which cannot have the knowledge we obtain from using it. "The reason the tape recorder does not know is that the information it receives neither generates nor sustains an appropriate belief."[59] It can be inferred from this that it is the capacity to form beliefs that "distinguishes genuine cognitive systems from such conduits of information as thermostats, voltmeters, and tape-recorders".[60]

As a preliminary part of the investigation, into whichsystems "qualify for cognitive attributes",[61] Dretske offers a division of intentionality into three levels. The first order of intentionality is described as contingent, being entirely dependent upon the interaction the system has with its immediate environment. The second order is nomic or natural, with the suggestion that the knowledge of this order is dependent upon the natural laws that hold empirically in the world. And the third order is analytic, for it is possible to have knowledge of *X* in the second order sense to form a belief about *X*, but not to know all the beliefs that are also synonymous with *X*.

A complete understanding of third order intentionality is only possible if we understand the term 'analytic'. An analytic sentence is tautologous, which is to say, it

contains no new information. It is of the form 'A is B', when {A, B} are semantically equivalent. For instance, the statement 'all bachelors are unmarried men' is analytic because the term 'bachelor' adds nothing to the meaning of the term 'unmarried man'. So both terms are interchangeable without any enhancement to the meaning of either term or any loss of truth-value in the overall statement.

Both knowledge and belief have, for Dretske, a very high order of intentionality. The beliefs we hold about *X* and the beliefs themselves are distinct even when their content remains interdependent. The simplest way of explaining this is that it is my having the belief about *X*, and therefore also an understanding of its semantic content, that makes it distinct from the beliefs about *X* per se. What has happened in my coming to hold a particular belief is that I have stripped away the superfluous information from my perceptual input and selected a specific piece of information, or concept, on which to concentrate. The information contained in the concept is then formed into conceptualised information from which the system is able to occupy a belief state.

2.8.1. Information content

Anything that has an information content is capable of exhibiting first order intentionality. A good example of this would be the visual field which carries information about the environment to the system. At this stage the system has a lot of information available to it and nothing in its perceptual field has been selected for attention. The information in this limited case is said, by Dretske, to be "analogue", which simply means that it is a continuous mass of information and not divided up for special attention.

2.8.2. Knowing

A more specific kind of analogue information is associated, by Dretske, with second order intentionality. If we use the visual field example again this is the system's narrowing of focus to the signals that it receives from a particular state of affairs in the environment. It is still analogue information because no one specific informational point has, as yet, been selected for attention by the system. In this sense epistemic states can,

according to Dretske, be attributed to the system that is capable of narrowing down its information content to this extent.

2.8.3. Believing

The third order of intentionality is related to what Dretske describes as "digital" information. This can be most usefully thought of as focusing on one specific object or event in the visual field. A specific signal in the visual field is selected and the information is digitalised, which means that the semantic content is reached and extracted. An important point to note is that it is not just the idea of focusing in on an object or state of affairs, but also the fact that one part of the signal is selected at the expense of the other signals or parts of a signal and also at the expense of the messenger carrying the signal.

Another useful way of thinking about this notion is to think of two sorts of watch. At a first glance an analogue watch gives a lot of very general information about the time, the face of the watch, its dial and so on, but if we look at the analogue watch using a microscope, (what Dretske would describe as second order intentionality), we will get much less information, the details of which will be more specific. When looking at a digital watch we get a very accurate account of the time, right down to the seconds; but it can offer no more than that. It gives only a very specific representation of the time and this Dretske would equate with third level intentionality.

Third order intentionality is equated with semantic content so that any system that is capable of this level of intentionality is also capable of understanding the nature of the object or event that it has focussed in on. Neither of the other two analogue levels are specific enough to offer an understanding of objects or events in their visual field. For Dretske the digital system would be capable of offering a representation of the specific object or event, whereas "in general analogue systems react to but do not represent"[62] the objects and states of affairs in their environments.

2.8.4. An example of focusing and selectivity

If we look for example, at the configuration of newsprint, we can see that both it and the visual experience of seeing it carry information. The visual experience cannot be classed as the semantic content since it is in analogue form and only that piece of information that is specially selected and 'digitalised' is the semantic content. The information carried by the visual experience is not digital; it needs a lot more cognitive processing before it can be raised to the level of a belief state.

If we consider a diagram of three concentric rings,which represents a newspaper, the newsprint and the information carried in the newsprint. The 'S' in the centre indicates that the whole diagram is a signal of incoming information from which information can be extracted. If you are looking unselectively you will see all the newspaper but no specific piece of information. A more selective look will still only focus on the newsprint but this time it will be on a particular piece of it. Finally, on close inspection an article or paragraph in the newspaper will be singled out and its semantic content sought. In the diagram below the semantic content is carried in the largest informational shell; the one in which all other information is embedded.

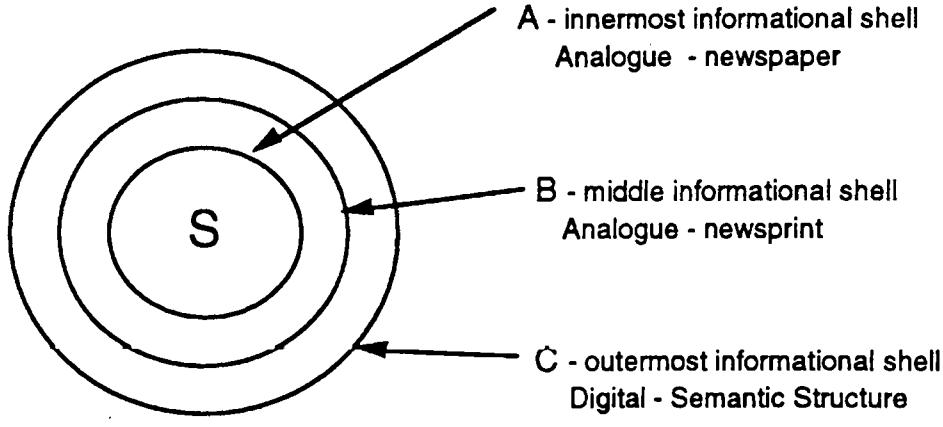

Figure 1

A system that is unable to read can still perceive the newsprint, so it is still able to receive the information even though it will not know what it means. However, the pattern of the newsprint alone will have no understanding of it and with a simple system there will be a coding deficiency happening where it is unable to move from the

perceptual form of the information to its cognitive form, i.e. there is an inability to completely digitalise the information that occurs in the senses in analogue form.

Dretske maintains that this 'inability' is why 'simple mechanical instruments' do not have access to the semantic structure of the information they receive. The instrument reads the information without forming any understanding of it. A voltmeter cannot completely digitalise the information because it is nested in other structures which the instrument is only 'seeing', and what it 'sees' is in analogue form.

In another diagram, Figure 2, it is possible to see more clearly the procedure that is involved in extracting the semantic content from the incoming information; a procedure we now recognise as 'digitalisation'. Again 'S' is the incoming signal, but the concentric rings now demonstrate the stripping away of irrelevant information to form a concept. The material that is relevant will depend upon the perceptions and mental attitude or disposition of the perceiver.

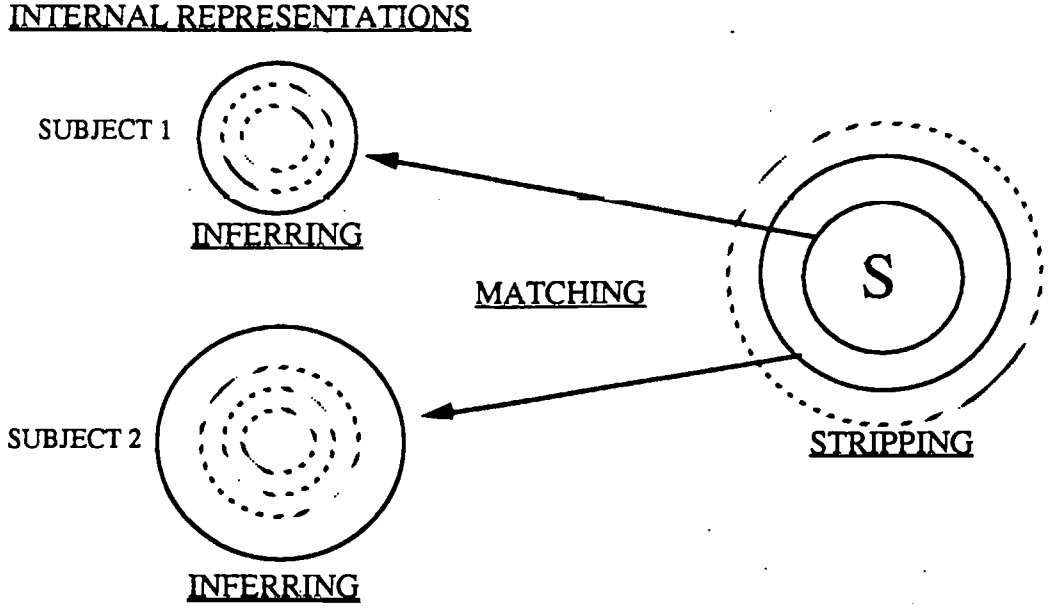

Figure 2

On the left hand side of the diagram the two sets of circles represent two different subjects each of whom has stripped incoming information away to reach one particular piece; a piece of information that is most important or particular to them. 'Subject 1' might be said to represent the typesetter of the newspaper; whereas 'Subject 2' might be the person reading the newspaper. The particular circles that represent specific 'chunks'

of information on the right have been matched with an internal representation of that same information. From the piece of information that they have chosen they can then infer what other circumstances will hold or be of interest and with this conceptualised information the system or subject is now able to occupy a belief state. (The circles with dashed lines represent pieces of irrelevant information.)

An instrument that is incapable of complete digitalisation is incapable of occupying higher-level intentional states. If a system can form a belief about its incoming information it can also intentionally alter its behaviour to suit the new beliefs that it has formed. However, only if the system is behaving in a rational manner will it take these new beliefs into consideration and this is what Dennett means when he talks of 'rational and predictable behaviour' that is exercised in relation to a system's assumed set of propositional attitudes.

In the instance of someone's holding a belief about some particular state of affairs we are not told the cause of their coming to hold that belief; all that is important, in Dretske's model, is that the individual's belief corresponds to the particular outermost informational shell and the semantic content of the overall structure. The belief carries only the information and says nothing about how it came to be there.

Dretske talks of a system having a 'plasticity' for extracting information. In essence what he means is the capacity for a system to ignore the message carrier so that it can concentrate on the information contained in the message. Perhaps a better word might be 'flexibility' because in everyday parlance we talk of being flexible in our decisions meaning that we are free to choose what we want to do. Few, if any, information-processing systems are capable of generating internal states from a remote source and ultimately being left with an information content which is also its semantic content. Indeed, for Dretske the flexibility to achieve this is only possible for those systems that can reach third level intentionality, being able to form beliefs from an analysis and synthesis of their incoming information.

The voltmeter can only carry information about the source by carrying information about the way in which the message is carried, i.e. the messenger; because of this the

pointer does not have as its semantic content the fact that the voltage is such and such. The information cannot be completely 'digitalised', therefore it cannot be a belief. And voltmeters and other 'simple mechanical instruments' can never occupy belief states.

2.8.5. Dretske - in conclusion

In conclusion it can be briefly said that Dretske's first and second orders of intentionality do not have any flexibility to extract information from incoming signals, and that his third order intentionality has the plasticity to extract information and produce the semantic content. So according to Dretske, the third level of intentionality is also to be thought of as the semantic content, and any system that is capable of this level of intentionality is also capable of understanding the nature of the object or event it has focussed on. Neither of the other two analogue levels are specific enough to have an understanding of the objects or events in their informational field.

The sorts of systems that Dretske equates these levels with is not very clear. Although, he gives us three distinct orders of intentionality based on the amount and extent to which information is processed, he only offers two categories of systems, the "simple information-processing mechanisms" and the "genuine cognitive systems".[63] In the first category he places 'dictaphones', 'television sets' and 'voltmeters', which, it should be noted, are all non-mental systems; and in the second category there are a mixture of mental and non-mental systems, for instance, "frogs, humans and perhaps some computers".[64]

So it seems that he, like McCarthy, Dennett, and Rosenschein is not above assigning mentalistic terms to systems that are more commonly described as non-mentalistic. His reasons for doing so are slightly different because they are set out in information-theoretic terms, but nevertheless they are there. He says, "If a system is to display genuine cognitive properties, it must assign a *sameness of output* to *differences of input*. In this respect, a genuine cognitive system must represent a *loss of information* between its input and its output."[65] With digitalisation information is lost, and this is why the television is not a genuine cognitive system for "it is incapable of

digitalizing the information passing through it".[66] A genuine cognitive system needs to be able to form beliefs and for this it needs the flexibility to extract or select information from a source which offers a variety of different messages. Such a system would also have to act in accordance with its newly formed beliefs or else we would have no sign that it actually had them.

If we now look at Searle's Chinese Room argument it will be possible to put the ascription of intentional states by these last three thinkers into a new perspective.

2.9. The Chinese Room: Intentionality, intrinsicality and semantics

In *Minds, brains and programs*[67] Searle sets out a distinction between weak and strong AI. The former is the view that AI programs are useful and powerful tools but nothing more; and the latter is the view that there are all kinds of possibilities for AI, for instance "that the programmed computer understands the stories and that the program in some sense explains human understanding".[68] Searle has no real disagreement with weak AI, but he puts forward 'The Chinese Room' argument as a challenge to the strong AI proposal that machines are capable of actually thinking.

The argument is probably already very well understood but, in case it is not, I will give an account of it again here. The argument is set up as a thought experiment and the idea is that you imagine yourself in a room which contains some baskets of Chinese symbols and a rule book for manipulating the symbols. On top of this you have to imagine that there are signs - which are questions - being passed into the room and you have to match them with other signs - answers to those questions - using a rule book. Having done this you pass the symbols back out of the room. The whole operation is done in a purely syntactic way without any understanding of the semantic content of the symbols. Having got proficient at the task of matching the symbols the people outside the room to whom you are passing them might well come to believe that you actually do understand Chinese because "your answers are indistinguishable from those of a native Chinese speaker".[69]

What Searle argues is that there is no way that you can learn or be said to know Chinese from the mere syntactic manipulation of Chinese symbols, thus the people outside the room would be mistaken in their view. You have no more understanding of Chinese when you leave the room than you did when you entered it. The point, he claims, is simple: "by virtue of implementing a formal computer program from the point of view of an outside observer, you behave exactly as if you understood Chinese, but all the same you don't understand a word of Chinese..... All that the computer has, as you have, is a formal program for manipulating uninterpreted Chinese symbols".[70] Neither you nor the computer can know what the symbols mean by just matching one pattern with another and passing the newly matched symbol out.

He argues further that "symbols and programs are purely abstract notions: they have no essential physical properties to define them and can be implemented in any physical medium whatever". [71] From this he concludes that because symbols have no physical properties that are theirs by dint of being a particular symbol, they can have no physical, causal properties. In a similar vein to Rosenschein's ISS approach, Searle says that the symbols in the program depend for their meaning on the programmer or designer. "Syntax by itself is neither constitutive of nor sufficient for semantics."[72] What is required for semantics is an understanding of the symbol's meaning. This might suggest that the symbols have their meaning intrinsic to them. This is not the case. For symbols to act as symbols they have to be about some thing, or things, in the world and the meaning has to be attributed to them from an external source. "The point is that there is a distinction between formal elements, which have no intrinsic meaning or content, and those phenomena that have intrinsic content."[73] The human mind is the only thing that can attribute meaning in this way so, Searle would argue, it must have some intrinsic mental contents (semantics).

Searle does not agree with the ascription of understanding to a computer. In fact he calls strong AI a "tin can and sealing wax theory"! For him the essence of having cognitive states, and in particular intentional states, is that they have an intrinsic or self-attributed semantic content. For example, our symbol manipulation in chess is what we

mean when we engage in a game of chess. This self-ascription is what Searle claims is missing in computational simulations. For in the normal running of a program the programmer need only know the syntactical set-up of the program.The program will work whether or not it knows its semantical content.

2.9.1. Does Searle weaken his foothold?

By answering the question, "Could a machine ever think?", with the answer "My own view is that *only* a machine could think, and indeed only very special kinds of machine, namely brains and machines that had the same causal powers as brains"[74], Searle does weaken his argument. On one hand he is adamant that machines cannot think and that they are just organised heaps of mechanical bits, whilst on the other hand he admits that human beings are machines as well! But Searle is a physicalist who argues that brains think and machines are physical entities in much the same way as brains. Thus, mechanical bits could think but not by virtue of an instantiated program, they would need some intrinsic mental content.

To a similar question, that of whether or not we could create a man-made machine that could think, he also answers "yes"; for he believes that it would certainly be possible if we were able to replicate all the physiological causes of consciousness. When he says "I regard this issues as up for grabs"[75] it is clear that he is not denying that one day machines, other than human ones, might be able to think and understand; so he does not flatly deny the possibility of a man-made thinking machine. Along the same lines he argues that a digital computer that could think could be created, because we are *eo ipso* digital computers.

However, he still answers the question, could something think by virtue of having an instantiated computer program, with a forceful "No". His reasons for doing so are that intentionality, and intentional behaviour, are biological processes and must depend upon biological phenomena, therefore they cannot be dependent upon only the formal processes that we find in a computer program. Up until now our computer programs can only simulate a thinking state and they can only do this because they are

programmed with the right things to do and the right order for doing them in. This programming makes their actions look like rational preconceived behaviour when it is not. Our ascription to the machine of thought and intentionality is what Searle would describe as "as-if" intentionality because the machine is non-mental.

Some time prior to Searle, Paul Ziff wrote a paper called *The Feelings of Robots.*[76] Many of the arguments in this paper closely resemble those put forward by Searle. For instance, Ziff argues that a robot may be able to calculate but not literally to reason in the way that we do. What we are continually doing with machines is over extending the metaphors we use until they lose their metaphorical sense and become meaningless in their new context.

Earlier still we find MacKay saying 'any test for mental or any other attributes to be satisfied by the observable activity of a human being can be passed by automata'; [77] nothing need be obviously wrong with the machine's performance, but it is still a performance and not the real thing. There is more to the intentionality of human beings than just their observable behaviour, yet we look only at the behaviour of a machine when we attribute the same intentionality. As Searle has said, "no simulation by itself ever constitutes duplication".[78]

Challenging the Chinese Room, (1):System semantics

But there are people who find fault with Searle's theory, for example, the 'systems reply' which has been made most forcefully by the Churchlands.

The systems reply states that "You don't understand Chinese but the whole room does. You are like a single neuron in the brain, and just as a single neuron by itself cannot understand but only contributes to the understanding of the whole system, you don't understand, but the whole system does".[79] So it is the room combined with everything inside the room that understands Chinese and not just the individual in the room. The idea is that whatever is doing the individual shuffling of the symbols is just like a single neuron in the brain, which by itself is unable to understand, but as one of

many it contributes to the understanding of the whole system. So the claim is that the system as a whole understands but the individual or single neuron does not.

Searle describes this challenge as a "daring move"[80] but refutes it on purely logical grounds. He reiterates that symbol shuffling itself does not mean that whatever is doing the shuffling has access to the meanings of the symbols. It will not make any difference to its inability to understand if we unite the 'shuffler' with its environment. If the symbol shuffler cannot understand Chinese then neither can the whole system it is contained in. Indeed Searle confronts this issue in the Chinese Gym argument; (a variant on the Chinese Room argument in which there is a "hall containing many monolingual, English speaking men").[81] The men behave like the nodes and synapses in the connectionist architecture of the brain. Again Searle argues that the Churchlands miss the point that was already made in the Chinese Room. They argue that a big enough Chinese gym would have higher-level mental features because of it size and complexity. But Searle opposes this by saying that any computation that can be done on a parallel machine can also be done on a serial machine. Thus if the individual in the Chinese room does not understand the language solely by carrying out the computations then neither can it understand when there are a whole host of them who do not understand. "You can't get semantically loaded thought contents from formal computations alone, whether they are done in serial or in parallel; that is why the Chinese room argument refutes strong AI in any form."[82]

An argument similar in many ways to the systems reply, though more architecturally explicit, is put forward by a great many proponents of connectionism. Connectionists argue that order can emerge out of complexity and if we have a complex enough parallel machine it will produce intelligence as an emergent property. The argument goes as follows "Consider a human being; most would say a human thinks, but at the lower level is that really the case? Does a single neuron "know" or have any self-awareness? I suggest the answer is no. So where does intelligent thought come from? It arises from the combined actions of many neurons together - it isn't programmed, but it is emergent."[83] The neurons can be said to be following rules (of

physics), in a similar way to the binary gates in a computer. In this sense the connectionist would ask why shouldn't computers think? We can take Searle's earlier refutation of the 'systems reply' to hold for this case also.

Sloman opposes the outcome of Searle's argument,[84] saying that "Computation is a purely syntactic, structural notion" whereas a "working computer goes beyond this. it understands its own machine code programs ...insofar as it systematically maps the bit patterns onto locations in its (virtual) memory and to actions which it can perform".[85] A computer can do a wealth of things, for instance it can compare, copy, modify and select from its incoming information if it has a set of instructions it can follow. However, Sloman does add that the world that the computer has access to is a "limited virtual" one, and that it is also constrained because it "does not have the full richness of human use of symbols"[86] to which it can refer.

Challenge (2):Intrinsic meaning

Another objection is raised by Harnad who questions whether or not meaning is intrinsic to a system. Searle, as we have seen above, wants to argue that intentionality is intrinsic to the human system, but not to a non-mental system like a computer. "Searle challenges.....that a symbol system capable of generating behavior indistinguishable from that of a person must have a mind",[87] but Harnad wants to argue that even though the manipulation of symbols in the Chinese room is based on shape and not on meaning, the fact that they are "systematically *interpretable*" means that the interpretation is intrinsic to the symbol system.

The symbols in a formal symbol system can only have meaning when they stand for things in the world. Such meaning cannot be intrinsic to that system since it is based on what the symbols mean for us. So the interpretation depends on the fact that "the symbols have meaning for *us*, in exactly the same way that the meanings of the symbols in a book are not intrinsic, but derive from the meanings in our heads".[88] The symbols in a book only have a meaning when we attribute one to them. So the

meanings in the symbol system are extrinsic and not a workable model for the kinds of meaning that both Harnad and Searle would say are intrinsic to us.

To get off the sort of "*merry-go-round*" we are on with the attribution of meaning to one symbol always depending on another symbol we have ultimately to ground the meaning in non-symbolic representations; which themselves have to have an intrinsic meaning if we are ever to get started. So Searle says that symbol meaning is intrinsic to us as part of our representational system, whereas Harnad argues that symbols come to have a meaning once they are used compositionally in meaningful syntactic ways, and that there is no such thing as intrinsic meaning - meaning is always something that is attributed. So the meaning of a symbol, for Harnad, is grounded in non-symbolic representations which are compositions of invariant features that are formed into meaningful and syntactical strings.

For Searle, the meaning is something that is intrinsic to a mental system and not to the symbols used by the system. It is this intrinsic mental content which makes mental systems more than just programmable computers. In a mental system the intentionality is, as it were, inbuilt by virtue of its being mental; whilst in a non-mental system meaning is attributed by an interpreter who is external to the system.

Challenge (3):Boden's response

On just this same point Margaret Boden argues that meaning is not intrinsic to the system, for, she argues, all meaning has to be attributed no matter what the physical make-up of the system might be. The attack she makes on Searle's "two pronged critique of computational psychology",[89] in the Chinese room argument, is both forthright and to the point; she says that the 'explanatory power' of his claims 'is illusory', "the biological analogies...are misleading, and the intuitions to which he appeals are unreliable".[90]

Broadly speaking Searle is claiming that computational theories in psychology are worthless. He says that only machines of a certain kind can think and that humans are, at a highly abstract level, digital machines but they are digital machines whose

substance is neuroprotein and not silicon and metal. Humans cannot be the simple instantiation of a computer program since the mere instantiation of a computer program cannot think, mean or understand, it can only shuffle uninterpreted patterns. A computer program for Searle has all the syntax and none of the semantics.

Boden says that Searle assumes that the computations of computer science are purely syntactic; that the computations "can be defined as *the formal manipulation of abstract symbols, by the application of formal rules*".[91] Intentionality cannot be explained in purely formalist terms since it gives no account of how the human mind employs the information derived from symbols it perceives. Searle, it would seem, equates understanding with intentionality.

Searle's second claim is that symbols only have meaning when they are embodied in something with 'the right causal powers' that can generate understanding or intentional behaviour. His argument is that the brain, and only the brain, has such causal powers , and that a computer does not. So, that machines are made up of silicon and metal is highly significant, for only something made of neuroprotein, like our brains, can have the requisite causal powers for understanding. Therefore, it is simply our biochemistry that makes us different from machines.

Boden's counter-attack questions the whole area of understanding in the machine and she offers two responses; the Robot reply and the English reply. The Robot reply states that if an automaton were to have all the input and output states of a human being, with limbs that could pick things up and with a compatible visual system, then it would be able to "demonstrably understand both restaurants and the natural language...used by people to communicate with it".[92]

Searle responds to this by saying that it just proves the point he was making all along that cognition requires causal relations with the world on top of the ability to manipulate symbols. He adds that whatever we add to the robot in the way of limbs and vision is still not adding intentionality or the cognitive capability of understanding. His response is 'personified' in the *Searle-in-the-robot* argument, where Searle places himself, figuratively speaking, inside the robot. In this case the incoming signals are no

longer pieces of paper handed in by Chinese speakers from outside, now they come through the robot's perceptual sensors. The limbs supply the capability for the machine output. Now Searle would argue that although the machine seems to be capable of whatever we are capable of it is in fact 'simply moving about as a result of its electrical wiring and its program' and it still cannot be said to understand.

Boden says that Searle makes an obvious mistake here, namely, that he considers that the robot is performing the functions of the brain then goes on to attribute full-blown intentionality to the brain. This, Boden argues further, is not something that computationalists do at all. "Computational psychology does not credit the brain with *seeing beansprouts* or *understanding English*: intentional states such as there are properties of people, not of brains."[93] Representations and mental processes are perceived as being part of the brain whereas propositional attitudes are ascribed to the whole person. And Boden argues that Searle has essentially missed the point that the computationalists are making. "In short, Searle's description of the robot's pseudo-brain (that is, of Searle-in-the-robot) as understanding English involves a category-mistake comparable to treating the brain as the bearer - as opposed to the causal basis - of intelligence."[94]

2.9.2. Computation is not just syntactic

Searle argues that the machine cannot reach the semantics of a formal symbol system but Boden maintains that "The inherent procedural consequences of any computer program give it a toehold in semantics, where the semantics in question is not denotational, but causal".[95] For any symbol to be part of a program it must have a meaning, and for the program to be the cause of other events it must be linked to some causal phenomena. We can find out what the symbol means by looking at the causal links to the external phenomena. The symbols in a program "do embody some minimal understanding",[96] even if we consider such an understanding to be "*so* minimal that this word should not be used at all".[97] Boden's conclusion is that "To view Searle-in-the-room as an instantiation of a computer program is not to say that he lacks all

understanding", for "computational psychology is not in principle incapable of explaining how meaning attaches to mental processes", for the embodiment of some 'minimal' understanding is not the same as understanding. In Boden's terms traffic lights, as symbols, embody understanding but they still do not understand.

2.9.3. Intentionality and biochemistry

Another criticism Boden makes of Searle is that if he is to maintain that it is our biochemistry that makes us different from machines, and that intentionality has its basis in the body's biochemistry, then why are we not able to define the products of intentionality in the way we are able to with the other bodily functions. She admits that, by its very nature, dealing with the brain is inherently more difficult that dealing with any of the other organs because of its inaccessibility and the essential part it plays in all our interactions. But she maintains, our notion of intentionality is more philosophical than biological, so, although Searle is probably correct about intentionality being dependent on the system's neuroprotein, he does not give us any information about why, or how, it does so. So, she argues that Searle is depending on intuition for his predictions and not on hard, empirical facts.

2.9.4. The humanist worries confronted

More generally Boden makes a number of points about the attribution of mental states to machines and the work of AI in general. She says that "The spectre of the mechanical mind haunts the lay consciousness because it appears to threaten deeply-held values and traditional beliefs".[98] Humanists could be said to possess such a 'lay consciousness' for in their philosophy, machines are incapable of carrying out truly purposive action; "being artificial or manufactured in origin, the machine's "thought" and "action" can be represented as meaningful or intelligent only by some ...appeal to the ends of the agent who made it".[99]

Persons who act to follow a goal that is not their own are often described as automatons. We call their behaviour mechanical. This, one might think, is sufficient justification for the humanist point of view, but no, their proposal also depends upon

whether of not the notions of purpose and meaning are intrinsic to the agent. Boden does not disagree with this position but she adds that the derivative use of a psychological term may be more of an advantage than the humanist thinks since computer analogies can help us to understand the workings of the human mind. Indeed many of the features of programmed computers are analogous to some extent to mental processes, and this has proved to be a useful tool. However, one of the big differences, and this is where the analogies begin to fall down, is that machines, as yet, do not approach their 'goals' from the position of having to wrestle with some internal conflict. Indeed, Boden seems to side, to some great extent, with the humanist arguments, but less on the controversial aspect of intrinsic meaning and more, because the technology to date does not even come close to the analogies we want to form with the structure and workings of the human brain.

Although Boden defends the use of analogous reasoning between minds and machines she does argue that "subjectivity, meaning, and purpose as currently understood can be attributed to artifacts only in the secondary sense, their justification ultimately deriving from the skill and interests of the artificer".[100] So there exist vast areas of mental activity where we can know only very little and then what we know is not through the construction of an analogous model, but through the capabilities of the designer. Indeed we may even discover that ultimately our notions of intentionality are largely indeterminate and thus nigh on impossible to program. In this instance, at least, Boden can be said to agree with Searle.

The number and content of our mental states, and in particular our epistemological states, mitigate against the possible application of all human thought in an inorganic system; "the epistemological issues involved are too obscure to allow one with a clear conscience to insist that *all* aspects of human thought could in principle be simulated by computational means".[101] So it looks like the magnitude and intricate nature of our mental states are, for Boden at any rate, probably the most compelling reasons for doubting that the postulation of human states is possible in an inorganic system.

This interest in the 'magnitude and intricate nature' of mental states, their *complexity*, suggests that the internal structure or architecture of the system will be important. This being the case I will now examine Sloman's work dealing with this area.

2.10. Intentionality depends on the complexity of internal architecture

Serial processors have been superseded by complex machines that are variably known as 'neural networks', 'parallel distributed processors'(PDP's) and 'connectionist nets'. In a seminar paper Sloman talks at length about complexity and its relation to the architecture of a system.[102] Essentially what he is saying is that the richness of a system's architecture is related to the complexity of that system.

In a section called "Against the Turing Test" he states that "behaviour is never conclusive evidence for mechanism...In particular, human-like behaviour does not *prove* the existence of coexisting, independently variable, causally interacting, more or less enduring, internal states, like beliefs, intentions, hopes, etc.". In this sense, at least, he does not seem to be very different from Wittgenstein, Searle, Dreyfus et al. Sloman goes as far as to say that, with a different physics than our own present one, all 'intelligent' behaviour could be presented in a giant "lookup table", although he gives no advice about how this would be achieved and what the new physics would consist of. The crux of the argument is that, because it is in principle possible to produce "intelligent-looking behaviour" in an "unintelligent mechanism" we can rule out the observation of behaviour as "a *defining criterion* for intelligence (consciousness, having beliefs, etc.)".[103]

So what we can be said to have in this schema is, at base, a single lookup table that cannot generate or cause any other states to occur, and, at top, a system that has the ability to generate behaviour. By comparison the lookup table will be a system with a simple structure, or architecture, whilst the top level system would have a complex architecture that gives it a greater capacity to act. The complex system is the one to which we usually ascribe mentalistic states; sometimes, at least in the work of Dretske,

Rosenschein and Sloman, such systems require such ascription because they are themselves, mental systems. However, mental terms are frequently ascribed to inorganic systems, and in the context of our interaction with such systems this seems to take place with a reasonable degree of success.

Sloman defends the use of such mentalistic vocabulary for describing inorganic organisms, but, he adds that a prerequisite of this type of ascription is that the systems are designed with a "richer (internal) architecture: more coexisting interacting states".[104] What he means by 'design' in this context is the architecture plus its mechanisms that can do all manner of causally related things; for instance: "creating, destroying, preserving, triggering, modifying, controlling, stopping, speeding up, slowing down, preventing, etc..." and all of these being directions for action between the component parts of the system as a whole.

In the discussion of 'design' Sloman talks specifically about creating interactive states between machine components. If we carry this back to the 'flesh and blood' example he talks of different mental states having different causal roles to play in the behaviour of the system. He gives two examples: "belief-like sub-states", which are environmentally produced and influenced; and "desire-like sub-states", which produce changes in the environment and which depend on the system's belief states.

2.10.1. Architecture and system capabilities

From this brief outline of the work it is possible to see that the 'design' and the architecture of the 'design space' are what Sloman perceives to be the most important aspects of the system if we are going to ascribe to it mental states. So with a different design space the system would be capable of different things and with a more complex design space, or architecture, the system would be capable of a greater number and wider variety of activities.

The thermostat is a simple case that is capable of only a limited number of procedures. It is possible, Sloman argues, to describe the thermostat as having a belief-like state that is varied by the environment (the curvature of its bi-metallic strip that is a

temperature sensor), and of having a desire-like state that is varied by the user (whatever temperature the user sets it at). However, he states that these are really "limiting cases" of the concepts and it is "silly" to argue over whether they are "real" beliefs and desires. I am not so sure that it is silly, nor do I accept that Sloman thinks it is either for he is still concerned about the architectural state, or states, under which we can accurately ascribe mental states to a system.

If the thermostat were architecturally able to detect the shivering or perspiration of the occupants of a room then its roles would be richer than its current on/off that we set and adjust according to our wishes. And, so, Sloman argues that "Different architectures support different mixtures of mind-like capabilities".[105] Simply a system's architecture is linked to its design space and different architectures will offer discontinuities in design space, meaning that the system would possess a different set of capabilities. Sloman describes the mental states of human-beings as requiring "VERY rich and complex architectures"[106], and we can infer from this that he accepts that human capabilities match the complexity of their architecture.

Indeed Sloman goes on to say that to look at the design space of a system we will need to take into consideration its requirements, that is, what it needs to sustain its existence, and the environment it occupies. Of its being able to have intentional states Sloman says that this requires that the system has "the ability to have representational states", this would enable it to distinguish between the intention to act now and ability to envisage future possible states and intend to act sometime then. The element of choice becomes important and the ability to choose between alternatives, that is acting now or acting later, will depend on the complexity of the inbuilt architecture of the system.

2.11. Concluding remarks

In the beginning of the chapter I looked at the problems associated with the theory of mental states and how these related to our definitions of mental acts and physical acts. Mental states, themselves, were seen to be important because it is through their

supposed existence, or the behaviour that is suggestive of their existence, that we are capable of ascribing them, or properties of them, to mental and non-mental systems.

Then I examined how the mind is related to its perceptual objects and the objects of its mental representations. It was noted that there is an element of reflexiveness in the relation that leads, in organic cases, at least, to a subjective or introspective, self-aware view of the system's interaction with the world. This can be compared with the inorganic case where reflexiveness may lead to adaptations or emergent properties, as in PDP's, but more often to predictable and generalised accounts of their relation to the world.

Within this area two views of the intentional relation were raised; the 'mentalese' or 'language of thought' of Fodor, and the 'intentional stance' of Dennett. Dennett proposed notional attitude psychology which was not dictated to by the nature of internal representations or where such representations would be located. Notional attitudes are part of the system's "notional world"; and the notional world is the world at a particular time and place that the system is best equipped, mentally, to deal with.

The essential role played by the nature of the system's experience of its world was seen to be important for its possession of intentional states. One of the most significant aspects of this was that for a fuller understanding of the behaviour of a system we need to be able to see the system in its behavioural context. This is especially so when we consider the ambiguity of our language and the problems involved in any sort of translation from behaviour to mental states. From the work in this area it seems that there are at least two possible conclusions, neither of which are completely satisfactory. The first is that we can depend upon behavioural capabilities for the ascription of mental states though as we have seen this is by no means guaranteed; and the second was that we could just resign ourselves to Searle's conclusion that there are mental systems and non-mental systems and 'never the twain shall meet'. As yet there seem to be no good reasons for accepting the fatalism of Searle's conclusion.

In what followed the discussion turned to intentional states and Dretske's division of them into the state where the system has only information, the state where the system

can be said to have epistemic states and lastly the state where the system can be said to be understanding, and holding beliefs. In this information-theoretic account, we saw the production of a hierarchy that gave different systems with different capabilities, different intentional states. Finally I examined Sloman's work that favours a swing towards architecturally related system capabilities, so that a machine with a rich architecture will have a wealth of, what might or might not be legitimately described as, mental states. It is this issue, and the preceding ones, that I will be confronting in the ensuing chapters.

Endnotes:

[1] Putnam, H (1988) *Representation and Reality*, MIT Press, p.1 ff.

[2] Descartes, R (1637) *Discourse on Method and the Meditations*, Discourse 5, Penguin Classics, Translated by F.E.Sutcliffe (1968)

[3] Casey, G 'Artificial Intelligence and Wittgenstein', *Philosophical Studies*, Vol.XXXI, (1988-1990), p.156-175

[4] Descartes, R. Ibid. p.73-74

[5] Ibid. p.74

[6] Ibid. p.74

[7] Bishop (1989) *Natural Agency - An essay on the causal theory of action*, Cambridge University Press, p.32

[8] Dennett, D. (1978) *Brainstorms. Philosophical Essays on Mind and Psychology*, Harvester Press, p.235

[9] Bishop - Ibid., p.32

[10] Ibid., p.39

[11] Ibid., p.39

[12] Putnam, H (1988) *Representations and Reality*, MIT Press p.1

[13] The term 'propositional attitude' was first adopted by Bertrand Russell in his "Theory of Descriptions"

[14] See Appendix 1 for a more lengthy exposition of the historical background of intentionality.

[15] Brand, M (1984) *Intending and Acting*, MIT Press

[16] Propositional attitudes possess a syntactic structure which can be understood, and because of this structure they are capable of being true or false. Their truth or falsity is independent of both the system in which the structure, or propositional attitude, is instantiated and the context in the presence of which it is had. It is only when propositional attitudes are expressed as statements in relation to a particular state of affairs that they become concrete bearers of truth or falsity. But I digress, for this is bringing us around to a notion of intentionality with an 's' that occurs in discussions of 'intensional' and 'extensional' language. (See also Appendix 2 for a further discussion of 'intensionality'.)

[17] Dennett, D (1987) *The Intentional Stance*, MIT Press, p.58

[18] Ibid. p.7

[19] Hobbes, T (1651) *Leviathan*, Penguin, Harmondsworth (published 1951), Pt 1 Ch 13, p.186

[20] 'Intrinsic intentionality' is usually meant as an aboutness that is different from other forms of intentionality because it is not derived from any other. Thought is an example of intrinsic intentionality, but linguistic sounds are derived from the aboutness of thought.

[21] The context of a sentence is referentially opaque if a term can be replaced within it by another term which refers to exactly the same thing but with a resulting change of truth-value for the sentence overall. Typically any propositional attitude statement is referentially opaque.

[22] Unless any particular language of thought was some sort species specific, in which case each individual from a particular species would share the same language of thought. The language of thought would be some sort of interpretational framework allowing each member of a species to understand its incoming information in a way that is essentially the same for all other members of that species.

[23] Husserl E (1913) *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy*, chp 27, p.51

[24] Ibid. p.53

[25] Ibid. p.52

[26] Ibid. p.53

[27] Lyons, W (1991) Intentionality and modern philosophical psychology - II. The return to representation, *Philosophical Psychology*, vol.4, no.1,1991 p.87

[28] Searle, J (1990) Consciousness, explanatory inversion, and cognitive science, *Behavioral and brain science*, 1990, p.587

[29] Ibid. p.587

[30] Ibid. p.587

[31] Ibid. p.587

[32] Ibid. p.588

[33] Ibid. p.588

[34] ibid. p.588

[35] Ibid. p.595

[36] ibid. p.595

[37] Ibid. p.595

[38] Ibid. p.587

[39] This passage of Searle's is very reminiscent of paragraph 293 of Wittgenstein's *Philosophical Investigations* where he talks of the "beetle" in the box. The "beetle" is representative of an aspect of private language, say of feeling pain. Only I can know what I mean by pain, I cannot point at my pain and add "That is what I mean by pain" for that would make no sense. Similarly, if I say I have a "beetle" in my box, and you are not allowed to look in my box, then you cannot be sure that our meaning of the word "beetle" is a shared, identical meaning. "The thing in the box has no place in the language game"; and so it is with aspectual shape for I can only perceive the objects I do from my point of view and I am ill-equipped to describe the aspectual relation between me and them.

[40] Searle, J (1990) Consciousness, explanatory inversion, and cognitive science, *Behavioral and brain science*, 1990, p.587

[40] Wittgenstein, L (1979) *Wittgenstein's Lectures Cambridge 1932-1935*, Basil Blackwell, p.113

[42] Ibid. p.113

[43] Ibid p.113

[44] Wittgenstein, L (1958) *Philosophical Investigations*, Basil Blackwell, paragraph 151

[45] McCarthy, J (1979) Ascribing Mental Properties to Machines - reprinted in Ringle, M *Philosophical Perspectives in Artificial Intelligence*, p.161

[46] Ibid. p.161

[47] Ibid. p.164

[48] Ibid. p.189

[49] Ibid. p.162

[50] Ibid. p.162

[51] Rosenschein, S (1985) Formal Theories of Knowledge in AI and Robotics, Technical Note 362 SRI International p.1

[52] Ibid. p.1

[53] Ibid. p.8

[54] Ibid. p.12

[55] Ibid. p.12

[56] Ibid. p.12

[57] Rosenschein, S (1987) The Synthesis of Digital Machines with Provable Epistemic Properties, Technical Note 412 SRI International

[58] Dretske, F (1981) *Knowledge & the Flow of Information*, Basil Blackwell, Chp.7 p.171

[59] Ibid. p.171

[60] Ibid. p.172

[61] Ibid. p.172

[62] Lyons, W (pers. comm.)

[63] Dretske, F (1981) *Knowledge & the Flow of Information*, Basil Blackwell, Chp.7 p.175

[64] Ibid. p.175

[65] Ibid. p.183

[66] Ibid. p.182

[67] Searle, J (1980) Minds, brains, and programs, *The Behavioral and Brain Sciences* (1980) 3, 417-457

[68] Ibid. p.418

[69] Searle, J (1984) *Minds, Brains& Science*, Pelican Books, p. 32.

[70] Ibid. p.32-33

[71] Searle, J (1990) Artificial Intelligence: A Debate, Scientific American, January 1990, Vol. 262 Number 1, p.21

[72] Ibid. p.21

[73] Ibid. p.21

[74] Searle, J (1980) Minds, brains, and programs, *The Behavioral and Brain Sciences* (1980) 3, p.424

[75] Searle, J (1990) Artificial Intelligence: A Debate, *Scientific American*, January 1990, Vol. 262 Number 1, p.21

[76] Ziff, P (1966) *Philosophic Turnings* - 'The Feelings of Robots' p.161-167

[77] MacKay, D.M. (1952) Mentality in Machines, *Aristotelian Society Supplement* XXVI, p.61-81

[78] Searle, J (1984) *Minds, Brains & Science*, Pelican Books, p.32.

[79] Searle, J (1990) Artificial Intelligence: A Debate, *Scientific American*, January 1990, Vol. 262 Number 1, p. 23

[80] Ibid. p.24

[81] Ibid. p.22

[82] Ibid. p.22

[83] Marsh, S (1991 - Nov) Listserv "Philos-l@uk.ac.liverpool" / Username: "spm@uk.ac.stir.cs"

[84] Sloman, A (1991) *Silicon Souls - Philosophical Foundations of Computing and AI*, (Notes for the AISB91 Tutorial), p.45

[85] Ibid. p.45

[86] Ibid. p.45

[87] Harnad, S (1990) The Symbol Grounding Problem, *Physica D 42*, 335 - 346, p.338

[88] Ibid. p.339

[89] Boden, M (editor) (1990) *The Philosophy of Artificial Intelligence* - 'Escaping from the Chinese Room', Oxford University Press, p.92

[90] Ibid. p.92

[91] Ibid. p.89

[92] Ibid. p.94

[93] Ibid. p.96

[94] Ibid. p.96

[95] Ibid. p.102

[96] Ibid. p.103

[97] Ibid. p.103

[98] Boden, M (1978) *Artificial Intelligence and Natural Man*, Harvester Press, Chp.14, p.418

[99] Ibid. p.420

[100] Ibid. p.425

[101] Ibid. p.444

[102] Sloman, A (1991) *Silicon Souls - Philosophical Foundations of Computing and AI*, (Notes for the AISB91 Tutorial)

[103] Ibid. p.12

[104] Ibid. P.12

[105] Ibid. p.17

[106] Ibid. P.17

3. Mental state ascription

3.1. Introduction

No firm distinction has yet been made between those systems that definitely possess mental states and those which do not. This being the case I will now discuss the reasons that underlie our ascription of mental states to other human beings, non-human animals and machines, after which I will begin to tackle the problem of whether or not the ascription can ever be justified.

The chapter complies with the following structure. I begin by stating the question that will be examined and clarified, then move on to reiterate the relevance of the question to this thesis and to theories of mind as a whole. Then I will pass on to some general arguments for why we ascribe mental states and a discussion of what, I argue, are the necessary criteria for the ascription of such states to a variety of human and non-human systems. As we reach the close of the chapter there will be a summary which will emphasise again the main points of my argument, and I will end by drawing each of these points together in a conclusory paragraph.

3.1.1. The question statement

In this section I will state and discuss the question of "What are the circumstances under which we ascribe mental states or intentionality to systems other than ourselves?". Such a question is important for many reasons which I will briefly recount here before setting out the central argument.

In the previous chapter I discussed several notions of the terms 'mental state' and 'intentionality' as used by a number of noted theorists in this area. Now, before we proceed to an examination of the issue of whether or not an inorganic system can have mental states, I will spend this chapter looking into some other related matters that have first to be taken into account. Initially I will examine why the ascription of mental attitudes and states has taken on such an important role in our world today. The second matter to take into account will be broken down into two parts; firstly, the ways in

which we recognise and identify different mental states, and secondly, the circumstances under which we feel satisfied enough with the criteria to attribute mental states to a system other than ourselves. Finally, I will look at what makes us, and here I am making an assumption, but no other system capable of attributing the phenomenon of mentality. Forthwith these will be known as the 'why', 'when' and 'how' of ascription.

3.2. The 'why' of mental ascription

The ascription of mental states and intentionality seems to be something that we do with greater frequency in our everyday existence. Most of the time these ascriptions are just implicit, which means that by our interactive behaviour we are tacitly ascribing particular mental states to other systems but at other times they are explicit, which is to say spoken. I believe that there are a number of related reasons for the rise in both implicit and explicit ascription and it is these reasons which I will now broach.

There have been many dramatic changes that have affected our environment over the last couple of centuries. In the eighteenth century there was the Industrial Revolution which began to provide machines capable of performing brute physical tasks, and so to irreversibly change the life and livelihood of mankind. Then at the end of the nineteenth and beginning of the twentieth centuries we were swept from a Classical or Newtonian view of the world to the non-determinism of quantum mechanics and Einstein's theories of relativity. Within a very small number of years our whole idea of physics had been turned on its head. So, with the scientific world trying to come to terms with events similar in scale to the Copernican Revolution of the sixteenth century, we find ourselves in this present century engulfed in the Information Revolution which is busy providing machines that will be able to replace the performance, by human beings, of brute mental tasks.

The fundamental characteristics of the Information Revolution have been the increased amount and availability of raw information and the tools to process it. Our communication systems have become so sophisticated that it is now possible to transmit

and receive messages in all sorts of ways; telephone, electronic mail and electronic file transfers to name but a few. It is even possible to work or converse with others using computers that are thousands of kilometres apart because of the rapid transference of data between multiple sites. Thus all the information we could ever require is now 'ready-to-hand' for we have both quantity and speed of retrieval at our command.

With all this freely available information the awareness we have of our lives, and the lives of others, has advanced so much that we are now capable of seeing ourselves in relation to a much greater context than was once available. We are no longer the inhabitants of a restricted social and geographical environment. Our increased information about our larger existence has made us aware of our world, and the universe within which our world is orbiting.

This revolutionary change has occurred for two fundamental reasons, the first is financial and the other intellectual. Both reasons can be seen positively as enrichments of the individual and his or her society. If we choose to look at them positively, they can be seen as rewards that have served to set up, sustain and forward a demand for more information and thus greater knowledge. As rewards they do, of course, encourage a greater interaction with our environment and one of the rewards of this is a richer awareness of our world. In this situation we are faced with an ever escalating pattern of behaviour; which dictates that with a greater awareness of our world there will be an increase in our desire for more information.

Of course, awareness and availability of information are not enough by themselves to make us intellectually or even financially richer; what is also necessary is an understanding of the received information, and an understanding that includes ourselves in relation to our world. It is understanding that is the crucial difference between the received or incoming information and the grasp of knowledge, and such an understanding itself depends upon at least a cursory notion of how to use the technology that conveys the information. It is not necessary to understand the internal structure and functioning of the machine, but an understanding of how to use the relevant technology is indispensable for an adequate exchange of information.

Thus we find that the information revolution has offered us new technology that enables the rapid conveyance of huge amounts of information to destinations throughout the world. Technology of this sort has been developed to do many of the jobs that used to be done by human beings. Then it is little wonder that, if the tasks that were once within the repertoire of only human systems that we know to possess mentality are now within the domain of those that are carried out by mechanical systems, we will ascribe mental states to non-human, non-mental systems.

If previously the performance of a task, or tasks, has been identified with human, mental systems, then a likely underlying presupposition might be that the minimum requirement for the successful performance of the task is that the system that carries it out is, at the very least, capable of occupying the mental states of which the human system is capable, or the equivalent of such states (whatever they may be). It is one way of making it easier to understand what sorts of characteristics would have to obtain for a system to be able to do some particular task, and from this knowledge it would be possible to anticipate the requirements of a system that would have to execute other similar or related tasks.

But it is not only a matter of our being able to know what are the necessary requirements of system A in circumstance B where it has to perform task C; another advantage of ascribing mental states to a system, even if that system has no mentality in the way that we have come to understand it, is that the ascription can be a very useful predictive tool. For instance, if we are able to predict to a fairly high degree of accuracy the actions of other systems, then it would mean that our interactions in the world can take on an order and determination that they would not otherwise possess.

With the ascription of a mental state being only that and not actually the ascription of complete mentality our interactions with any other system (even with another human being) can only ever be carried out more shrewdly and with better informed judgements. For unless the other system is identical in all ways to me, and with a personal history that is exactly the same as mine, its actions can never be determined by me in the way that I can determine my own actions. If it had all these properties it

would be feasible to argue that the 'impostor' was in fact me for it could be no-one else. If it were possible to ascribe a full mentality, or active mental life, it would have to include aspects such as a freedom of will and again, because the system could act in whatever way it pleased, we would have come full circle to find ourselves once again only capable of informed, but not accurate, prediction.

That we have the capability to predict at least a proportion of the prospective behaviour of other systems means we are no longer interacting in such a haphazard way within a random world. Being able to predict action means that we can also adapt our own behaviour with respect to what we expect another system will do and this adaptability gives us an increased chance of survival. If we had no such capability, that is, we were to have no idea of the temporal or causal connections between past, present and future actions, we would have to learn each event for the first time each time it were to take place. Inevitably this would have serious consequences for the survival of the human species.

Now, the question of whether we are right to make such ascriptions is not, as yet, the issue; for we do ascribe states to other systems and there is a rationale behind this action. Dennett and others would argue that it is perfectly reasonable, in fact, perhaps even inevitable, that we will ascribe mental states to anything that exhibits 'human-like' behaviour; indeed it is necessary in some sense to do so because it enables a much more sophisticated interaction with such systems, which in turn enriches our understanding of whatever information is passing between us.

People from the humanist[1] schools of thought would argue to the contrary saying that ascription of mental states to non-human, inorganic systems can only possibly entail misunderstanding since we are saying of a machine or artifact that it, for example 'knows x' or 'believes x', when it is not in fact capable of such complex activity. They would argue that the requisite mentality is missing from such a system and they would say that Dennett, McCarthy and others are guilty of using language that already has an application in one particular context in a quite different context where its use and

meaning are perhaps subtly, but significantly altered. I will now look briefly at the differences that context can make to the use of language.

The language we use to describe the actions of organic systems has been developed to suit a context in which there are certain organisms, for example, human beings, that behave in certain ways that are strongly suggestive of their possessing a variety of sometimes complex and sometimes not so complex mental states. When we use the same language to describe the putative mental states of inorganic systems that are also inanimate, such as a teddy bear, we are told that this is mistaken attribution and that we are guilty of anthropomorphism. However, for some people it seems to be relatively unproblematic to describe a system that is inorganic, but with moving parts, as being in possession of mental states.

This seems to suggest that the difference is in the possession of the moving parts, but if that were so then describing a Jack-in-the-box as having a desire to surprise would be quite a natural and acceptable thing to do. No, the difference must lie in something more than a thing's just being capable of movement, for if that were all that was required for the possession of mental states mentality would not be an issue at all.

The something 'more' that the is required by the system is to be capable of exhibiting appropriate behaviour. What is meant by 'appropriate behaviour' is that which would lead the observer (and ascriber) to say that the system is in possession of some mental state that corresponds with the behaviour. If we look at the example of *trust* the distinction between mere animation and appropriate behaviour can be seen to stand out more clearly.

As human beings we have the ability to trust each other when the occasion warrants and for the most part our trusting someone depends upon their exhibiting behaviour that we are able to interpret as being that which is trustworthy. However, it is true that we extend our trust to some inorganic things thus enabling us to say of a car that we trust its brakes, its steering and so on. Thus it is possible for us to drive with a feeling of greater security than we would otherwise have.

Both of these judgements are about our having a relation of trust with another entity and about being able to predict, with some degree of certainty, future events, yet they are different in very significant, but straightforward, ways. For instance, the trust we have in another human being is inextricably bound up with our prior knowledge of that person's character and our interpretation of their behaviour. In one sense my trust is about the physical behaviour of the individual, but in another sense it is about what I believe to be going on inside that person's head. On the other hand, trusting that my brakes will not fail is just a belief about the physical world and the physical world states that entail. It depends upon the beliefs I have about my world and not about the attribution of mental states to the car or its brakes. It follows that both examples are about physical events but the former is also about the mental states that precede and accompany the physical events when performed by an individual with a mental life.

Another sign of the difference, but this time a purely linguistic one is this: I say that my trust in a friend can be betrayed if my friend lets me down in some way; yet I do not say of my brakes that they have betrayed me, (unless I am being melodramatic), for it would be a peculiar misapplication of language since the term 'betrayal' is reserved for use with things that we consider to be morally culpable and thus responsible for their actions. The worst my brakes can do is fail, but they cannot betray me as a trusted friend might.

Of course, it is true that in the event of an accident we might say "I trusted these brakes" or "I blame the brakes", but this is just because of our tendency to ascribe intentionality and mental states in an effort to explain circumstances that we might not fully understand. If the brakes fail they do so out of a physical deficiency, but if my friend betrays me she does so out of choice. The difference is in the fact that only one of them is capable of making a decision about its actions and the other is entirely dependent upon whatever physical states of affairs hold at that time.

In a similar way, if we were capable of constructing an algorithm from which we could predict the behaviour of even the most complex computer then, in terms of trust, all we would ever have is the sort of trust that we can have in the brakes of our car. The

sort of trust we have in other people is a different kettle of fish that is more like having faith in a god, that is, the sort of belief that can be firm, indeed even unshakable, even when the evidence for our holding the belief is scanty. An algorithm from which we could predict the future behaviour of other people would not be a formalisation of trust, rather, it would reduce the need for us to trust at all.

I will now turn to the reasons why the evidence upon which we base our attribution of mental states is, at best, inadequate. A bare outline of my argument is as follows: our ability to ascribe mental states depends upon two things, namely; our ability to use and understand language, and two, our apprehension of the complexity of the system with which we are dealing. (These will be discussed at in section 3.4 and then at greater length in chapter four.) We have already seen that animation is not enough and that appropriate human 'mental state' behaviour is necessary if we are going to even consider the possibility of imputing mental states in an inorganic system. I would like to look at what counts as 'appropriate' behaviour, and when is it possible for us to recognise and identify such behaviour as such.

3.3. The 'when' of mental state ascription

To reiterate, we are looking for what leads us to attribute mental states to other entities, and so far we have settled that the only evidence we can go on is the perceivable behaviour of the system to which we are attributing the states and so in this instance the 'when' describes the state of affairs under which we feel justified in our ascription of mental states, and the justification can only be when the system behaves in accordance with the paradigm case 'as-though' it understood, believed, knew, wished or whatever.

The problem with evidence and finding 'appropriate' behaviour is that we have to first of all establish behaviour with which it can successfully be compared. Once we have this model we can say 'Yes, this behaviour fits with our model' or 'No, this behaviour does not accord'.[2] So that any behaviour other than the appropriate one would not, as a result, be in accordance with the system having that mental state.

The most suitable starting point in the search for appropriate behaviour is to look at the entities to which we already ascribe mental states and look at how they behave in some given circumstances. The most obvious entities are those that we know to possess mentality because it is with them that the attribution of mental states will most readily take place. So we will look to human beings, for it is with them that we can start to construct a paradigm of behaviour.

3.3.1. Recognition and identification of 'appropriate' behaviour

The discussion that follows will be about the sorts of evidence we have for the occurrence of mental states, and because of this it will be about mental states in general and not any particular example such as, 'believing', 'knowing', 'hoping' and so on.

As already mentioned there are two rules to follow to recognise that a human being has a particular mental state. The first of these is to look at the individual's behaviour, for in the majority of cases (there will always be exceptions) each different mental state will be discernible by a different behaviour or repertoire of behaviours. The second rule is to look for spoken verification of the mental state that the person is claiming to occupy. This type of behaviour is characterised by the use of the first person propositional attitude, for example, "I believe" or "I wish". I will argue that neither of these are completely reliable methods for an accurate determination of the mental states of any system, but, until something more dependable comes along, these are the only guides we have.

When wanting to know if someone believes something we look for evidence of that belief. Here are two examples, one of a belief in some physical, and therefore observable, aspect of our world, and the other a belief for which there is no physical object in the world, occupying a point in space and time, to which the belief relates.

The first example is of Mary who puts up her umbrella, thus giving us an indication that she believes she needs some form of protection from the elements. If she goes on to put on some boots we will have another piece of evidence, from which we can extract the information that Mary believes she needs shelter from wet weather. Mary

may also offer us a verbal verification for our thinking that she has a certain belief or set of beliefs; for instance she may say "I believe it is raining and if I don't have protective clothing I will get wet". Thus we have two sources of proof for the attribution of a belief state to Mary, we have her physical behaviour and her spoken behaviour. We have also our own empirical corroboration to back up Mary's verbal and visual evidence. So if we see it is raining, and Mary acts in accordance with there being bad weather, then we are more likely to attribute to her the belief that it is raining.

The second example is of someone holding a belief in something superphysical, a divine being or a state of Nirvana, perhaps. For us to attribute the mental state of belief in this case we can only rely on the acts of the individual, for example, they may chant mantras and wear saffron robes, or they may attend a particular church service on every appropriate occasion. On top of this they might try to convert us by telling us about their own personal epiphanies or how much better their lives have been with a faith in something spiritual. In this, the abstract example, no empirical experience of our own could ever corroborate the mental state of the other individual for his or her experiences are of a personal or subjective nature.

So in our first example we have the physical and spoken behavioural evidence of the person to whom we are ascribing mental states, plus our own visual back-up to add credence, or indeed a refutation, to their story. In the second case we have only the behavioural evidence of the other person for there is nothing that we can perceive that could add or subtract from their being in a particular state of belief. An interesting point about the second example is that my ascription of a mental state does not only rely upon the careful interpretation of the individual's behaviour but a necessary aspect of my ascription is an examination of their behaviour within a specific social situation. It is this contextual dependency that makes my identification of their mental states more reliable; for in a whole context I am more likely to recognise if I am being deceived, say for example, that the person to whom I am ascribing mental states is an actor who is taking part in a theatrical production.

I have shown that for the ascription of mental states we first need a model with which we can compare the behaviour of a second system to which we might, given the right circumstances, make an ascription of mentality. In this instance the paradigm case is that of human activity, for human beings are already complex systems that not only possess mentality but also the ability to communicate their experiences of their individual mental states. By using their example to show how we can compare, recognise and identify the greatest likelihood of the occurrence of particular mental states, I have been able to demonstrate that ascription depends upon one of three possibilities. These possibilities are: firstly, to make an ascription of mentality solely on the basis of physical behaviour; or secondly, to ascribe mental states on the basis of behaviour that also has a linguistic back-up, that is, the person saying they 'believe x' or 'desire y'; or thirdly, to make an ascription of mentality on the basis of behaviour that is linguistically reinforced by the person but in addition to that to have the ascribers own, perhaps, corroborating experience of the system's world at the time that the ascription might take place.

Having demonstrated that we are able to make ascriptions of mentality to systems other than ourselves on the basis of exhibited behaviour, I will now move on to the question of how we actually undertake such ascriptions.

3.4. The 'how' of mental ascription

In this context 'how' will be used to describe the physical manifestations behind the action of making an ascription. For example, how we actually ascribe a mental state of belief or unhappiness. I shall be arguing that the act of ascription can be made in either one, or both, of the following two ways; directly through the use of language, or indirectly through our interactive behaviour with another system. Thus in this section I will discuss these two acts of ascription.

3.4.1. Language - linguistic ascription

For me to be capable of fully understanding both the physical and verbal behaviour exhibited by another system I too have to be a language user; for I need to be able to

understand what is being said by another system as well as be able to use language to express what I think it 'knows', 'understands', 'believes' and so on. So by being a language user that already manifests mental states I am capable of attributing mental states to systems other than myself. This does not preclude the ascription of mental states using a non-verbal form, but it does suggest that behavioural ascription could not be made as clearly as linguistic ascription.

Being a language user means that I have all the procedures for verifying the mentality of others at my disposal. For instance, I can observe the behaviour of another human being, I am able to understand when it uses propositional attitude statements to express its state of mind, and I am able to apply the value of my own experience as a testimony to its stated frame of mind. Were I interacting with a non-human system, incapable of using natural language, it would only be usefully possible for me to compare its behaviour with that of other more complex systems. Only its non-linguistic, or physical, behaviour is perceptible; and it cannot describe its own internal states reflexively in a way that human beings can. Even someone who is mute can make gestures, such as a sweep of the arm meaning 'all of this', that shows that he or she sees themselves in relation to the larger context of their world; a computer is not capable of this sort of self-consciousness behaviour.

It seems then that being a natural language user has its advantages for it allows the user to assign meaning to things in its world and thus to interpret ever changing states of affairs and to act on them by attributing mental states where applicable thus enabling a more thorough and sophisticated interaction with its environment. There is also the advantage that information can be passed quickly between users of the same language, for example, the same social group who have shared meanings and uses of words. I would like to now put forward an argument to show that the acquisition and use of language are important for our being able to ascribe mental states.

Language acquisition and use are important for ascription

In our everyday use of language its acquisition does not play a big role, indeed it could be successfully argued that how we came to acquire language plays no part at all in day to day conversation. To demonstrate this I will offer an example that portrays a physical interaction between me and my world. If I wish to cross a river and I make enquiries about a means of doing so, the origins of the language that I use are unimportant. What is important is that I am understood; how I came to know the word 'bridge' or to make grammatical use of the preposition 'across' is immaterial. Similarly, I am not interested in how my means of crossing the river came to be there, unless perhaps to question its safety, but even that is more a matter of its physical structure here and now rather than a reflection on the skills of the bridge builder.

Thus I would argue that when talking of physical things in the world, such as a bridge across the river or a game of bridge, we need make no recourse to how our language came about. For a proper understanding one needs only to have learnt the relevant language and be able to use it in the appropriate circumstances so we can understand each other and make ourselves understood. If we use language in inappropriate circumstances it will sound like nonsense. The appropriate circumstances are what Wittgenstein would describe in the *Philosophical Investigations* as the right "language game".

We start learning how to use language, or play "language games", from a very early age without any formal or theoretical instruction. We may learn the names of objects by ostensive definition or in association with other things, but we only learn their application through interactive use in society. Throughout our lives we continue to learn new words, with different applications and configurations, and because of this how we acquire language becomes gradually less and less important compared with the manner in which it is used. Both the acquisition and the use of language have to take place in a social setting, but of both of them only the language use remains socially important,

since it is only through continued shared use of words and phrases that we can ever know that we are using them in an acceptable way.

Up to now I have been concerned with the acquisition and use of language in relation to talking about everyday states of affairs in our worlds. I have claimed that how we acquire language is irrelevant for its successful continued use, but I have added that how we use language remains important, if only because it allows us to have shared meanings, thus enabling the process of communication to take place. According to Wittgenstein a "private language" is of no use for communication because no-one could ever know precisely what any other person means. A shared use results in a shared meaning.

In the slightly different context of ascribing intentional states there are new problems to be met. No longer are we confronted by ordinary language which is used to describe a physical world, now we are faced with trying to offer a description of mental states that lie, by their very nature, undisclosed to us. We are back to the philosophical problem of other minds but in this case we are concerned only with when and how we should ascribe the capabilities of minds to systems other than ourselves. If we talk about the ascription of 'capabilities' to minds rather than the ascription of actual mental states the problem of other minds becomes one of physical functionality, which is to say what the system is able to do, rather than whether it possesses intentionality and has mental states that are comparable to the ones had by me.

Earlier I said that we attribute mental or intentional states to other systems if they seem *as-though* they know, desire, wish, and so on. So the ascription of a form of mentality can be seen to depend entirely upon creating analogies with other behaviour that we associate with particular mental dispositions. Even from an examination of my own behaviour I can see that there are occasions when I behave in a 'belief-like' or 'want-like' way. If I then extrapolate from these and examine the behaviour of other systems it might be possible to pin down which essential characteristics occur in both my behaviour and that of the other system. I will have created an analogy between the two sets of behaviour.

If I am going to use language to ascribe an intentional state to something that I am doing, thinking or whatever then I am going to have to first of all be capable of discerning my own mental states from my behaviour, and then I will have to be able to compare my behaviour with that of other systems when the same intentional language has been used. To do this will require a great deal of understanding of both the behaviour that is taking place, and the language that is being used to describe the behaviour.

My social use of language is very important if I am to have a complete understanding of both the language that the other system uses[3] and the language I use to describe my own mental events. However, in the case of mental ascription the way my language was acquired is also of great importance. For if, as I have argued, language is acquired and used through social interaction, and the ascription of mental states and intentionality is first made by analogy with my own mental states, of which I have a first hand knowledge, and subsequently with other systems with which I interact socially, then the way language is acquired and subsequently used will matter a great deal. It is from our initial acquisition of language that we learn and build up the framework of linguistic behaviour, the propositional attitude elements of which can then be characterised by the exhibition of appropriate 'mental state' behaviour.

If we acquire a language by analogy with things, situations and states of affairs in the world and if that language is in continual use then the way it came into being stops mattering. However, with the language that we use to describe mental states its acquisition is by analogy with that behaviour which we consider to be the most appropriate 'mental state behaviour', so how the language was originally acquired is important if the analogy is to be upheld.

Such mental state behaviour is necessarily human for two reasons, firstly, because we know humans possess mentality, and secondly, because they are the only systems with the capability to describe these mental states. The analogy set up at the acquisition stage of learning must always be significant, even if we are not always aware of it,

since it is on this basis that we can attribute mental states. The first analogy is important even if the subsequent contexts alter and we learn some new criteria.

The most notable difference between the ascription of mentality and the physical use of language is that the behavioural analogies of mental attitude ascription, by which we learn to recognise the likely occurrence of a mental state, only alter very little throughout our lives and our continued use of language. Their alteration is restricted because they are the outward physical signs of the manifestation of an inward non-physical mental state and we have no other way of recognising them.

Quite simply the recognition and identification of mental state behaviour is different from the recognition and identification of physical objects because of the disparity in the quantity and quality of the received information. With concrete objects there is always much more informational input; for instance, if I am learning how to identify a pint of beer I will have the taste of the beer, its smell, the colour of the liquid, the size of the receptacle and many other things to go on. For the recognition and ascription of a mental state we have much less information to go on but what we have must remain closely akin to its original form if recognition is to continue.

Before moving on to the next section of the chapter to show that the ascription of mental states can also be non-linguistic, I will give a summary of the argument that I have set out above. Then I will draw the attention of the reader to some examples of attribution through linguistic interaction to a human system and an inorganic system.

The acquisition and use of language requires a social environment, an environment in which we can learn to use language properly, that is, the way our society does. How we acquire the terms to describe physical objects in our environment does not continue to matter, but how we use those terms does. How we acquire mental state language is by inward reflection and outward use. If I feel sad and have no language to describe it, I will nevertheless still feel sad; and the people with whom I interact would behave with me 'as-if' I feel sad. With the facility of language it is easier to express my sadness but it is still difficult to compare it with the sadness of another person.

The acquisition of my 'sadness' language will have been from my own experience of my mental states when I feel a particular way. My use of language to express my sadness will always be a social use that involves interaction with other human beings who can also feel and express (perhaps using slightly different language) their sadness. How I acquire my experience of sadness and the language I learn to accompany it will always be important for it is that acquisition that dictates how I judge my own later experiences of sadness and the sadness I experience, second-hand, in the hearts and minds of others.

When dealing with non-human organic systems, such as cats or beavers, the ascription of mental states is most definitely done on a basis of non-linguistic interaction, however, when dealing with human beings the attribution of mentality made by the human observer can be verbal as well as behavioural. The human capability to ascribe mental states both behaviourally and linguistically means that it has a distinct advantage over other systems that are not capable of using natural language. The advantage of being a language user is three-fold: firstly it means that the description of our own mental states is more exact; secondly, we can be understood by other language users; and thirdly, our ascription of the mental states of other systems will be more accurate.

However, it is not just the ability to use a language that matters, it is also the fact that there was a need for a more competent form of communication there in the first place followed by a continued use and adaptation of the language in our social environment. A recent argument states that "it is not the fact that man can speak which counts so much as the fact that he has something to say".[4] That we have something to say is a result of the elaborate nature of our society and our complex interactions with one another within our society. The complexity of our interactions is made more so because we are are trying to define the nature of abstract, 'in-the-head' entities; the mental states of ourselves and possibly other systems.

3.4.2. The ascription of mental states using language

Now to two examples of the ascription of mental states through linguistic interaction. The first is of a computer that asks for the replacement of a floppy disk. Often when trying to run an application on a computer a message will appear on the screen asking that the user replace a specific disk. The disk will be one with the application on it and the computer is unable to run the application without the disk. The message is usually something like "Please insert the disk entitled 'Word4'". To all intents and purposes the system can be attributed with a 'need' for the disk, or it might even be said to 'want' the disk.

The 'want' or 'need' can easily be compared with a human need for something; and with this example in mind, perhaps the need for a pencil and some paper before a letter can be written. It is certainly true that human beings can start the whole operation of letter writing in their heads and later transfer it to paper but the actual letter writing cannot be started unless the implements are there to be used. The 'implements' needed by the person are comparable to the 'application' needed by the computer.

In this example I have deliberately kept the comparative needs superficial, so that the need for a floppy disk by the computer runs no deeper (cognitively) than the need by the human being for writing equipment. What are more difficult to compare are the deeper emotional needs of a human being for affection and security, for nothing similar exists in the computer environment. Even an animal such as a cat can show a need for affection and warmth by pushing its head against your hand until you stroke it or curling up on your knee when you are reading a book, but no computer has yet been developed that needs to be encouraged or praised when it has correctly transferred a file or transferred a text from Latex to Word 5. Indeed it would take a vigorous stretch of the imagination to mistake the computer's superficial request for the much deeper wants and needs that can be expressed by a human being or some animals.

It should also be remembered that the computer is intended to be 'user-friendly' so it is meant to emulate the kind of polite human-like behaviour that is most likely to get a

positive response. The language the computer uses is programmed into it so that it can be more easily understood and our use of intentional language[5] is just a way of describing the actions of the computer in a way that is likely to be understood, and more importantly responded to, by competent users of the same, in this case natural, language.

The second example to be used is that of someone running to catch a bus. There can be no doubt in the mind of the observer that the actions displayed by the other person of looking round, trying to cross the road hurriedly, running along near the edge of the footpath, continually glancing behind them to watch the approach of the bus and shouting "Stop", are the behaviours that correspond most accurately with someone having a desire to catch a bus. If I was to go through a process of gross analogy with my own behaviour every time I thought about ascribing a mental state to something I might say to myself, "If I were doing all the things this person is doing now what would be my state of mind?". At least using this technique I am able to compare my own mental activity with what I assume to be theirs.

It is true that unless we are acquainted with the person we are unlikely to know why they hold such a desire, but our ascription to them of the wish to be in time for the bus does not depend on background reasons, only on their behaviour at the time. It is certainly true that if the person had not shouted to the driver to "Stop" I would still have been able to infer from their other behaviours that they wanted to catch the bus. It is to this non-linguistic behaviour and ascription that I wish to turn to now.

3.5. The story so far

This section is a summary which will be followed by an explanation of what exactly is meant by the 'non-linguistic ascription of mental states'. Up unto now I have talked about why, when and how we set about ascribing mental states to other systems and here they now are in précised form.

Why do we ascribe mental states? We ascribe mental states for two reasons, the first is that we are interacting with inorganic systems that can perform mental-like tasks,

and, the second reason is that it is a useful predictive tool that facilitates interaction and communication between human beings and what are perceived to be 'intelligent' systems.

When do we make ascriptions of mental or intentional states? Mental states are ascribed to systems that behave 'as-though' they understood, wished, hoped and so on. Since I can never know for sure what is going on inside another system all that is left to me is to base my ascriptions on their perceivable behaviour. There are three levels of verification for a system being in a mental state; they are, purely behavioural, behavioural with the system offering a linguistic back-up, and finally, behavioural with a linguistic back-up and the corroborating experience of the ascriber.

How do we physically ascribe mental states to other systems? Ascription can be either linguistic or behavioural. Which is only to say that I can attribute a state to something by saying "X wishes that Y" or "Mary believes it is raining", and in each of these the attribution is made using language; alternatively, I can attribute a state of mind to something by adapting my behaviour to fit in with what mental state I perceive it to have, and ascription of this latter type is non-linguistic.

The distinction can also be made with respect to implicit and explicit representations. For instance, I would not say of a machine that it 'believes Y' but I might behave implicitly to it 'as-though' it does. The machine or the cat might behave as-though it has a particular goal in sight. It behaves in totality, that is, the system altogether can behave in such a way but there is nowhere within the system that we can say 'Ah, there is the representation of its goal'. The goal-directedness is a function of the system as whole and no one particular aspect of it. In sum we can say of the machine, and perhaps all animals except human beings, that the appropriate goal-directed behaviour is accompanied only by an implicit representation.

The human system seems, at the moment anyway, to be the only system that can have explicit representations. An explicit representation can take a number of forms, spoken, as when I say "The moon is made of green cheese", or thought, as when I think but do not say out loud that "The love of money is the root of all evil"[6], or even

physically drawn but not verbalised, such as a painting of my favourite walk. The capabilities of human beings are enhanced by the fact that they can have such explicit representations. It follows then that any system that could move from having only implicit representations to having explicit ones would necessarily be endowed with a different set of capabilities, and not least of these would be the fact that it would have graduated from 'as-though' mentality to having real mental states.

3.6. Apprehension - ascription need not be linguistic

The non-linguistic ascription of internal characteristics takes the form of a brute physical enquiry. A sort of 'poking and fiddling' approach which can probably be best described with the help of examples. The first example will be of inorganic systems and the second will be of non-verbal communication with an organic, but non-human system.

Say, for example, I am given two machines to use and the only way I can find anything out about them is to fiddle and poke at them. If one is a thermostat and the other is a video recorder my interaction with each of them will begin to take a distinctly different form. There are fewer buttons, if any, on a thermostat and only one switch with an on/off position. The video recorder, on the other hand, has many buttons, lights and switches all of which cause me to behave with it in a more circumspect way. On the basis of these gross physical differences alone my interactions with the different systems begin to develop and become separate. If I go further and open both systems up to see what is inside this will only serve to reinforce my by now distinct approach to each of them.

On a superficial interaction alone I will have come to the decision that the video recorder is a much more complex system that is capable of a great many things that the relatively basic thermostat is not. By my behaviour alone I will have ascribed to them quite disparate internal states. The thermostat is only capable of low-level states, such as being able to process information about the temperature of the room, whereas the video recorder can be programmed to record different programmes at different times on

different channels. My behaviour with the video recorder will probably reflect the instinctual feeling I have about this being quite a 'clever' piece of machinery that can seemingly understand what I ask it to do.

A not dissimilar interaction is the non-linguistic interaction we have with non-human organic systems such as pigs, cats and beavers. The biggest difference between these systems and non-linguistic inorganic systems is that the former have mental lives about which we know very little. We do not, for instance, know if a cat or mongoose is capable of seeing itself in relation to its world or whether its responses are just a matter of the organisation of innate DNA structures that dictate the drives for fight, flight, food and reproduction. With inorganic systems we can be fairly certain that the ones we have to date are just *finite state machines,*[7] that is, machines with completely determinable states, that are incapable of introspection or exercising free-will.

It may be feasible to take an inorganic system to pieces to examine its internal design and machine states but it is not a feasible procedure for our investigation of animal or human mental states. Although it is possible, with some experiments, to look more closely at mental states such as excitation and stimulus-response, it is not possible to discover anything about what the system believes, hopes or feels sad about since we have not yet achieved the dizzy heights of being able to recognise in patterns of brain behaviour the mental states which might belong to them.

We attribute mental states to animals in a very similar manner to the way in which we ascribe mentality to humans; we base it on 'as-though' behaviour. If a cat is hungry and it has grown up being fed canned food that is taken out of a particular cupboard, then it is odds on that the cat will have learned by association that rubbing itself against the cupboard and miaowing will elicit the desired response of a willing human with a can-opener and a tin of cat food. The cat behaves as-though it knows where the cat food is and as-though it knows how to ask to be fed. If we are willing to attribute mental states to inorganic systems because of their exhibited behaviour then there seems no reason why we should not attribute mental states to camels, cats, and crocodiles on the basis of their display of 'knowing' behaviour.

3.6.1. The implications of 'as-though' ascription

With the help of these examples it has been shown that there are often good reasons for ascribing mentality to systems that exhibit behaviour which suggest that they are occupying a particular mental state. The ascription in these examples is behavioural but non-linguistic because the systems being dealt with have no use of natural language. There are two possible implications of this type of ascription, namely: one, that the mentality is in the head of the beholder and not in the system being apprehended; and two, that mentality is something that is also out there and by interacting with other systems we get glimpses of it.

The first of these conclusions runs headlong into solipsism because it denies mentality to anything else, including other human beings, and permits mentality only for me in my world. There can be no meeting of minds in this world for my world is all that there is since all I have available to me are my own perceptions. A weaker version of this view might be that it can be accepted that mentality exists in other human beings on the basis that if it exists for me, and they are like me physically, there is a very good probability that it exists for them. In this sense it is possible to create our human analogies of mental state behaviour for it is possible to imagine what it would be like to be another human being. What becomes difficult is imagining what it would be like to be another entity like a cat or an aardvark, and even more difficulty is encountered in imagining what it would be like to be a computer, for we have no experience with which to compare their being. Which is only to say that in some sense I know what it is to be an organic entity but not what it is like to be an inorganic entity.

To accept the second implication is to accept that mentality can exist in any other system that behaves as-though it has mental states. It is a position adopted by realists who accept that there must be mental activity in other things because we are able to interact with them in complex ways. For a machine to behave as-though it understood, and for it to have an implicit representation of the question being asked, is still a form of mentality even if only a simulated form. So it is not worth discarding realism out of

hand. The biggest drawback for realism is that it, like solipsism, ceases to enquire about the nature of mentality once its position has been adopted.

A more attractive position is in the middle ground where it is possible to ascribe mentality to other systems on the basis of a created analogy with my own experiences. Their experiences may never be identical to mine but they, that is my mental states and their internal states which we can only glimpse in their behaviour, have similarities and overlaps that are by no means insignificant. The most we can say for now is that our access to other kinds of mentality is limited and can be attained only gradually, and we may eventually discover that the relationship between my mentality and that of other non-human systems (animals and machines) is asymptotic.[8] Thus we might find that many, but not all, of our mental states can be shared with other types of systems. One particularly problematic state is that of self-consciousness, as the following diagram illustrates.

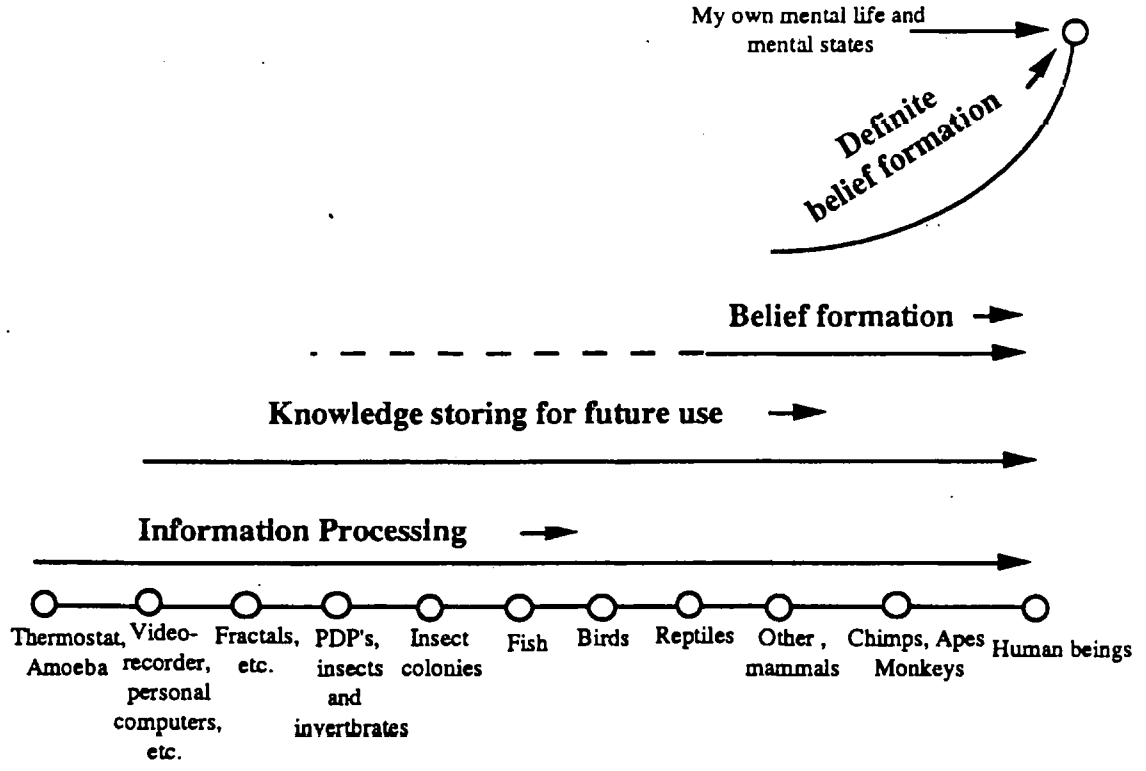

Figure 3

In this diagram the dashed lines indicate the problems that exist in trying to differentiate between one sort of mental state and another, for example, being able to say where a knowledge state ends and a beliefs state starts is by no means an easy task.[9] The diagram is described as being *asymptotic* because the lines of 'Information processing', 'Knowledge storing' and 'Belief formation' respectively, all run in the same direction, above the systems that are capable of performing at each level, however, none of them actually reach and follow the curved line of 'Definite belief formation' which indicates the beliefs of which I can have first hand experience, which are, my own beliefs. Notably none of the other systems can ever 'reach' and thus have direct experience of my own self-conscious intentional states.

3.7. What role has vanity in our reluctance to ascribe mental states?

Throughout this chapter it has become apparent that the ascription of mental states is done on the basis of whether the exhibited behaviour can be described as human-like, and if it can, in what sense and to what degree is it human-like. A new term, 'as-though', has been employed to express the similarity of the behaviour but also to try to deflect the conflict that surrounds the subject of non-human systems that behave 'as-though' they are in possession of human mental states.

I would like to put forward an argument to propose that initially the conflict arises out of a sort of arrogance that allows human beings to think of themselves as having an unrivalled intelligence and a monopoly on the possession of high-level mental states. An intelligence such as it is that walks hand-in-hand with the knowledge that we are subjective systems capable of introspection, the contemplation of abstract concepts and the ability to ascribe mental states to other things and of these mental states we imagine that if they are in existence, we will be able to draw an analogy between them and our own. So, we ascribe mentality to something that behaves like us because of a belief that says if the other system is capable of behaving like us then it must have mental states just like ours else how could it behave in the manner in which it does.

An arrogance of this sort can be seen in modern astrophysics in the case of the "Anthropic Principle" which states that "there will only be observers to look at the

Universe in rather specially selected universes".[10] The implication of this is that our cosmology interacts so intricately and with such accuracy that without it existing in just this way the human being could never have come into existence. We are part of a "specially selected universe". Thus it can be concluded that human beings did not just happen by accident but are instead part of some larger well-fashioned plan in which it was ordained that our existence would come about in order for us to be able to understand the universe and its order. What the Anthropic Principle does not say is that the other implication is this: if things had not happened just as they have the lack of 'plan' would not have mattered to us because we would not be here to reflect upon it.[11]

Here we have human beings that perceive themselves as being comparable to, but not identical with, any other system which is a dangerously exalted position from where it is easy to topple.

3.8. Social and observational criteria in ascription

Before I move on to the summary of this chapter I would like to say a word or two about the question of intrinsic meaning and how it relates to the problem of ascribing mental states. When dealing with symbols and the manipulation of symbols the idea of intrinsic meaning is contrasted with that of attributed meaning; the former suggesting that a symbol has a particular meaning as an essential part of its being that symbol, and the latter, that all symbols have a meaning assigned to them by the users of the symbols.

Both intentionality and the ascription of intentionality can be expressed in the form of propositional attitudes, and for this the system must be capable of using language. To use a language it must first be learnt, and to use it to form intentional statements the system must be capable of incorporating its language into its everyday life. The element of uncertainty implicit in the phrase 'everyday life' suggests that the system must be capable in some sense of analysing and synthesizing the language it possesses in order to form new expressions that will describe new states of affairs.

So a fundamental requirement for a system that can have belief states and intentionality is that it be capable of manipulating symbols. This ability means that a language user is able to juggle symbols, assign new meanings to symbols and create new expressions from novel configurations of symbols, and because of this it has a tremendous advantage over any other non-linguistic system. One example of this advantage is that the language user can interact with other similar systems, sharing and conveying information that is mutually beneficial.

Through social interaction with other systems of similar linguistic potential the human system has become a language user. It is both capable of using language to convey useful information and of creating language to make this process possible. Thus, it follows, that it must be able to manipulate symbols and strings of symbols to express new meanings and also attribute new meanings to symbols that have an established use, as, for example, when writing a code.

We are, in effect, *human symbol processors* that can manipulate formal symbol systems, but we are distinct from the purely formal symbol manipulators because we are capable of the assignment of meaning to whole sets of symbols and perhaps creative meaning or language us when writing works of fictional literature or song lyrics. A good example is of the invention of a whole new alphabet where each letter will have been assigned a meaning distinct from any of the other symbols, each consequent conjunction of letters will have a complex meaning and each configuration of these conjunctions into words and phrases will have a separate set of meanings still.

As language development and use is a product of the social interaction of linguistic peers the meaning of its constituent parts, whether singly or in conjunction, will be dependent upon its culture. Similarly, language can only be relevant when used in the culture in which it originates and is currently in use. A crude example to show the importance of this relevancy is that of English speakers who go abroad and try to be understood using their own language, they do, sometimes without noticing, raise their voices and speak more slowly in a vain attempt to be understood, not realising there are also cultural barriers acting to inhibit language comprehension.

With this cultural development of language it can be argued that it is not possible for symbols to have an intrinsic meaning for meanings are 'put into' our heads by linguistic interactions. Meaning becomes a social phenomenon that makes it possible to speak to ourselves internally in a sort of 'internal socialization', but the meanings are not in our heads alone; they are a composition of what is in our heads and in our worlds. In just this way the meanings we ascribe to things are a reflection in our heads of the reality in the world. As a social construction the essential elements of any language will have their meanings attributed by the parts of society that use that language. I, for example, would understand very little of a knitting pattern, but people who knit regularly could converse happily and successfully, each understanding the other and being understood.

To summarise: that symbols can have intrinsic meaning becomes very doubtful if we accept that the symbols and strings of symbols that go to make up a language are created and adapted by a society for the use of that society. The symbol meanings are attributed by the users, and it is only through using a language in the environment in which it was developed that we ever come to know that a meaning is shared.

The implication of this to our current problem is that it comes full circle to demonstrate our reliance on social or observational criteria for the ascription of mental states to other systems. To say of something that it understands is to use a social phenomenon, language, to describe a social phenomenon, behaviour. Both types of social phenomenon can only be fully understood when the language and culture are also understood. If the meaning of a symbol was intrinsic then that symbol could be applied cross-culturally with no loss of, or mistaken, comprehension.

The meaning of a language is something that is attributed by the users of that language; it is based on social criteria and does not mysteriously exist somewhere in the head. In just the same way mental states are attributed, they do not have a physical existence and are socially defined and dependent.

3.9. Main summary and conclusion

My discussion began with how widespread and frequent has become our ascription of intentionality and mental states to non-human systems. It was proposed that this was due to our increased interaction with a largely technological environment, and that we ascribe a variety of mental states to these systems because they perform tasks that were previously only imaginable within the domain of human beings.

Non-human, inorganic systems are now capable of doing things that were once only done by human beings who are capable of wide ranging high-level mental activities. As we are the only system we know to possess mental states we ascribe mental states by forming an analogy with ourselves. We are human beings that have a mental life with intentionality and mental states which we can describe using language for other language users to understand. We have found, perhaps through trial and error, that it is better to treat a system that behaves 'as-though' it has a particular mental state as though it really has the state for it assists our interaction with it and permits us to narrow down and predict its possible future states.[12] If all we have is the behavioural criteria that suggests that a system is occupying a mental state then it would be more sensible to go on this information than ignore it in the hope of something better turning up.

A big advantage that we as human systems have is that we are capable of using language which enables us to describe our own intentional states and inner mental life and to express our ascription of mental states to other systems. The way we acquire and use language was shown to be important for the attribution of mental states. We acquire language through analogy with objects and states of affairs in our world and we only learn to use language properly through shared use in a linguistic society. Then when we come to ascribe mentality to something this too must be done by analogy since we have no access to actual mental states other than our own. The analogy we look to has to be with other things we recognise as being part of our linguistic interaction with our world.

However, as was shown, language is not the only way of ascribing mentality for it is possible to ascribe it through behavioural interaction alone. Everything depends upon my behavioural interaction with the other system in a way that suggests what its capabilities might be, thus implying whether or not I think it capable of only 'processing' the information, 'understanding' the information or forming 'beliefs' about it.

From this discussion it can be concluded that if we are to achieve a good understanding of the possible mental states of another human system there are a couple of conditions that would have first to be met. They are, firstly, that we, and the other system, would need to be users of a shared and mutually comprehensible language so that verification can be given of the mentality of the system to be ascribed; and secondly, that the ascription of mental states would also need to be made using the same language. Where there can be no linguistic corroboration, for example, a human being who is mute or a non-human animal, our understanding of the mental states can only be partial. But even here it would be possible for us, as human beings, to put ourselves in the place of another human being and imagine how he or she must feel even though they cannot tell us. In this sense then, it is possible to have a better understanding of a human being who is without language than an animal who is also non-linguistic.

Understanding why a machine acts in a particular way is easier to discover than any kind of understanding we can ever hope to achieve of the behaviour of a cat or another person. This is because it is possible to know in total the internal structure of the machine and to know what state it is occupying at any one time. Thus it is possible to know of what the machine is, and is not, capable. In this way, it is extremely doubtful that of a thermostat one would ever want to say, "It believes the room is too hot". What we might be more inclined to say is "The thermostat has processed the information correctly and the relative curvature of its bi-metallic strip has caused the heating system to be switched off". The second statement is without question a much more accurate representation of what has gone on inside the thermostat.

It must always be borne in mind that the events that take place inside the 'minds' of other systems are only available to us through shared behaviours which can be either, or both, linguistic and physical. Eventually we have to say that something does or does not possess a mental state and it is only a matter of accepting as evidence, though not incontrovertible, what we perceive before us. Only the other system can ever know that it occupies a particular frame of mind or mental state. The implication of all this is, of course, that even if it were possible to endow an inorganic system with mental states only it would ever know for sure that it had such states because only it would have direct knowledge of them. Only the individual system can speak the "language game" that describes its own private mental states; so it is just as Wittgenstein says "If a lion could talk, we would not understand him".[13]

---

1 Also known as the "carbonist" school because they maintain that the only system capable of possessing mentality is a system that is made up of carbon atoms.

2 The words do not have to be spoken or even said 'into oneself', for when we recognise appropriate behaviour we may just act in accordance with the other system being in possession of a particular mental state. In this way our comparison of the observed behaviour with our model has been implicit.

3 In this case the language refers to that used by another human being or the programming language that is created and implemented by a human designer for human/computer interaction.

4 McCrone, John (1992) *"Avoiding the Freudian Slip"*, Computing, pg.35 20th February 1992

5 Intentional language is more often seen in messages like "I can't find the dictionary. If you find it for me this time I promise to remember where it is in future." that appear when you want to use a Spellchecker in a word-processing application.

6 1 Timothy, chapter 6 : verse 10.

7 A full explanation of Finite State Machines will be given in chapter five where I discuss four different kinds of machine and their capabilities to recognise increasingly more complex grammars.

8 More will be said about this asymptotic relationship in later chapters, and in particular in chapter 7, the conclusion.

9 This problem will be discussed in much greater detail in chapters five, six and seven.

10 Longair, M. (1989) 'The new astrophysics', p.201, taken from *The New Physics*, Paul Davies (Ed) Cambridge University Press

[11] Similar sorts of discrepancies can be brought to bear against the teleological arguments for God.

[12] Any implication that a system is in a particular mental state is still only an implication, it is never without doubt.

[13] Wittgenstein, L (1958) *Philosophical Investigations*, Basil Blackwell, p.223.

4. Complexity

4.1. Introduction

This chapter will be structured in the following way. As in the previous chapter I will begin with a statement of the problem area and the specific question that is to be confronted. Following this I will look at the notion of complexity as three sub-divided issues.[1] The three categories of complexity are, (i) the architectural complexity of the system, (ii) the complexity of the action or behaviour of a system, and (iii) the complexity of the relationship between the system and its environment. They are subtle distinctions which will be drawn together again in section 4.3. when I will look at how a system's architectural complexity can be related to the complexity of the capabilities that the system can perform. I will also attempt to show that the second and third categories necessarily collapse into one since no behaviour can be exhibited without there being some relationship between the system doing the behaving and the environment in which the behaviour takes place. To finish the chapter I will give an account of how these notions of complexity arose in chapter three and how the issues of ascription in that earlier chapter relate to the broader notions of complexity that will have been unfolded here.

4.1.1. A statement of the problem

In this chapter I will examine the question of how machines can be distinguished and stratified by means, and in terms, of their complexity. I am proposing that the notion of complexity can have many different interpretations and that this, in itself, makes the task of distinguishing between systems a lot more difficult. However, in an attempt to confront these difficulties I state the problem as follows: given a specific task or competence, what is the minimum degree of complexity that a system would require to accomplish it?

If one were to examine the question of consciousness, one should begin by looking at which systems we already accept as possessing consciousness. We would then need to ask the question of why these systems qualify but others do not and one approach to this might be to look at what other capabilities are manifested by the system and then examine the relationship between these other capabilities and the presence of consciousness. It might then be suggested that the capabilities that accompany consciousness in one system will be most likely to occur in other conscious systems. Then, in an effort to see which other systems possess consciousness, I could apply the now specified criteria of 'requisite capabilities' and if they are present I might extrapolate that consciousness is also manifest in the system.

Because the thesis as a whole deals with the attribution of mentality, and whether or not it is justifiable to ascribe mental states to inorganic systems, the tasks and competencies that I will look at will be mental ones such as understanding, knowing and believing. I will be assuming that for organic systems to have mental states such as these they must also have consciousness, for without it their mentality would be inactive or redundant. In chapter five I will illustrate and explain Dretske's hierarchy of mental states that will show that some mental states are of a higher order than others, and because Dretske proposes that there are such levels only complex systems with many capabilities can reach what he cites as the highest order mental states.

With this borne in mind a distinction will be maintained between possessing consciousness and possessing self-consciousness, for, as yet we have evidence to show that only the human system is capable of being self-conscious and thus manifesting higher-order mental states, and this, of course, complies with our notion of the human system as a remarkably complex one. I will now look more closely at three categories according to which complexity can be defined.

4.2. Three categories of complexity

4.2.1. Complexity of architecture

It is possible to demonstrate that there are three quite different notions of complexity. I shall set them out as distinct notions and then go on to show how they overlap with one another. This section will be concluded with a discussion of which notion, or notions, can best be used to define the sort of complexity that is at issue here.

The first notion is that of the complexity of the system, as in "here is a system that has a complex internal structure". Forthwith I shall call this "architectural complexity". In this instance the primary concern is with the internal design and structure of the system and with nothing that is external to it. Because I am concerned here with organic and inorganic systems, and because many people would quibble with the use of the term 'design' for the internal structure of an organic system, I shall, henceforth, talk of the 'architecture' of the system and this shall refer to whatever is internal for both organic and inorganic systems.

When trying to establish the differences that exist in architectural complexity between different kinds of system it becomes apparent that it is not as easy as one might at first expect. For example, if we think of the internal organs that I have in common with my cat then we find that there are very few differences, for it too has a heart, kidneys, liver, lungs and a brain. It is true that the cat's organs may function in slightly different ways to mine. For example, its heart beats faster per minute than mine and its cooling system is different for it cools down by panting, thus allowing water to evaporate from its tongue. I on the other hand perspire so that there is a surface covering of water on my body which evaporates and causes my temperature to decrease.

There is the gross physical difference of size but this leads us into another dead end for, if I were to conclude that because the cat has a much smaller brain it is less capable than a human being, then I would have also to plead that because the elephant's brain is

larger it is more capable than a human being, and this is quite evidently not the case. However, one ratio that is important is that of brain size in relation to the overall mass of the animal. A good comparison can be made by looking at the size of the human brain in relation to the size of its body and then looking at the size of a dinosaur brain in relation to its size. It is quite easy to see from this why dinosaurs have been considered to be stupid for they have a tiny brain that seems to bear no relation to the immensity of their body. Penrose writes that the part of the human brain that human beings are "proudest" of is the cerebrum -"for that is not only the largest part of the human brain, but it is also larger, in its proportion of the brain as a whole, in *man* than in other animals".[2]

Of course, when we consider the difference between the architecture of a machine and the architecture of a human being we can immediately see that there are very obvious physical anomalies. Fundamentally the machine is made of different material, it is silicon and metal, whereas human beings and other organic systems are carbon based skin and bone structures. However, the external nature of a system is not always a sure indication of its internal architecture for some machines can carry out much more complicated tasks than, say, a hare, so no hard and fast distinction can be drawn to show relative complexity between systems on the basis of their physical constituents. Thus, in this instance at least, the skin versus metal distinction can be passed over, even though it is the source of a fundamental property distinction between organic and inorganic systems.

The internal architecture of a machine can be seen to be in many ways dissimilar to that of any living system. A machine has no heart for it has no need for a supply of blood, nor has it a brain that needs oxygen for nourishment. What it does have is a supply of electricity that feeds it in a way that might be thought of as analogous to the blood supply in any animal body, and it has an elaborate arrangement of wires, silicon chips and circuit boards that take the 'nourishment' of the electrical charge and carry out the computational operations it has been programmed to perform.

It must be said that computers are not trying to emulate, in every possible way, the functions of every mammalian organ; but what they are trying to model are some of the many functions of the brain. To make the model look more brain-like, and perhaps act more brain-like, the internal architecture of many computational machines has been developed to resemble a neural network, bringing with it the advent of parallel, instead of serial, processing. One advantage that parallel processing is often assumed to have is that it can run and complete tasks in much less time than is needed by a serial computer. However, this is not always the case for, by their nature, some computations are better if processed serially. The quick rule to follow is to look at the task and see if it can be divided into subtasks that can each be carried out independently of any of the others. If it can, each of the subtasks can then be processed concurrently and, as a result, the overall processing time will be speeded up.

With gross physical differences in architecture being ruled out as indications of relative complexity the distinction might turn out to be more subtle in nature. An example of a subtle difference can be seen in the primate family, of which anthropoid apes, such as man, monkeys and chimpanzees, are all members. The physical differences between the different members of the primate family are found to be negligible; we are roughly the same size, our limbs, body and head are arranged in the same fashion, and we all have opposable thumbs with which we can lift, hold and use tools. The architecture of our brains, as in all other mammals, is also roughly similar, and the size, shape and structure of the monkey brain has led to numerous studies of its behaviour being carried out so that it is possible to see how closely man and other apes are related.

Monkeys have been observed to be capable of many things and not least of these is the capability to work through quite complex tasks by acting out all the necessary behaviours, often with the use of tools that might have been previously created. An example of this is trying to reach food by breaking into a termites nest which needs three separate types of stick and three distinct stages of activity. The first stage is to use a heavy stick to break the shell of the nest, next to use another stick to poke a hole in

the nest, and finally another stick is used with which the termites can be drawn out. Indeed, going through this complex repertoire of activity it might be considered that the monkey has an explicit representation of the goal that he/she is attempting to achieve. Thus our distinction between man and other apes seems destined to lie in the subtlety of a two percent difference in our DNA structures, for ninety-eight percent of our DNA is identical.

Of course, a geneticist might argue that this relatively small amount of difference in the DNA structure is really quite substantial and not the subtle distinction that I am suggesting it is; but I would contend that I am discussing the likelihood of the complexity of architecture being placed in direct relation to the possession, by the system, of different kinds of mental states, and this is not genetics. So, the amount of identical DNA that humans and chimpanzees share might turn out to be a good indication that apes and chimpanzees are only slightly less complex in structure than man but still capable of manifesting high-level mental states such as knowing or understanding, even forming beliefs and possessing self-consciousness, - as we have seen Penrose has claimed.[3]

One of the major problems that is encountered when trying to compare the complexity of the internal architectures of different kinds of system is that we are often attempting to compare two unlike things. For instance, the inside of a computer or a kettle is very different from the inside of a bat or a sea-cucumber, so trying to establish some relationship between their comparative levels of architectural complexity is not really all that feasible. However, if we look at the architecture of any of the marine coelenterates (jellyfish), and compare their architecture with that of the dog next door, we will discover some similarities that enable us to construct a more realistic comparison. Both organisms need food to live, and both have digestive chemicals that aid the break-down of their food; and both organisms need to take in oxygen to live, although they do so in very different ways. Both organisms have cells and neurones but in the dog they are thousands of times greater in number and in less primitive formations than in the jelly fish.

Similarly, it is possible to compare the architecture of a thermostat and a video recorder for they are essentially constructed out of the same stuff and just by looking inside, as we did in chapter three, we can see that a thermostat has only a few parts and a very simple design. As a result of its architectural simplicity the thermostat would be easier to replicate. This is not the case with a video recorder.

Another set of examples that should not be forgotten are those of systems that have a simple internal architecture but a complex array of behaviours. In the computational world complex patterns can be formed from simple equations as seen in the Julia and Mandelbrot sets. Whilst in the animal world we need only think of the ant or the bee for they are both capable of behaviours that seem to go far beyond what one would suppose possible from their limited structure. Both types of organism have evolved complex social structures in which different members of the group play different roles.

Leaf-cutting ants of South America have huge underground nests and are capable, by working together, to bring down a tree, remove all the leaves, shoots and stems and carry it back in tiny pieces to their nest. Once there they chew the pieces of tree to form a compost and feed off the fruiting bodies that are produced by the compost. Another example are the tree ants of Southeast Asia that sew leaves together to construct a nest. This is made possible by a sort of competition where one group of ants hold leaf edges together with their jaws and feet and another group on the inside of the leaf sew them together. The sewing material is produced by them bringing larvae to the site and squeezing them to produce silk. The ants doing the sewing move the living larvae across the leaf junction until the leaves are finally joined.

Thus it would seem that whether the distinctions to be made are gross or subtle the internal architecture by itself can lead us to few conclusions about the overall nature and complexity of the system. So I shall move on to consider another possibility which is that it is only possible to demonstrate the difference in architectural complexity by an examination of the capabilities that an inorganic system is designed to accomplish, or that an organic system can be seen, by its nature to, possess.

4.2.2. Complexity of capabilities or behaviour

In this section I will look at the second category of complexity which is the complexity of the system's action or behaviour; for example, "here is a system that can do complex things". Forthwith this will be called "behavioural complexity". In this category only the external behaviour is important and not what the architecture of the system itself is like. As in the example above, of the monkey and the termites' nest, the monkey was seen to be capable of planning for the action needed to carry out the task of acquiring food and even using three separate tools to enable it to do so; I am concerned now only with what a system can be seen to be capable of doing. The relative complexity of these behaviours will then be taken as a reflection of the complexity of the system that is capable of carrying them out.

Again if we look at the example of the thermostat it has a limited repertoire of actions that are primarily dictated to it by the particular construction of its binary mechanism and bi-metallic strip. It cannot act in any other way than it does because that would necessitate a different structure and as a result it would be a different machine altogether. A sewing machine, on the other hand, is capable of carrying out a greater number of different functions than a thermostat so it can be correctly assumed that it has a more complex mechanism. Still more numerous and varied are the functions of which a basic computer is capable and it is justifiable for this reason to say of the computer that its architecture is still more complex than either of the other two. However, they are all only capable of processing the information which they have been designed to receive, and for this reason my thermostat cannot sew a patch on my trousers, my sewing machine cannot transfer files from one directory to another and change the respective format of the documents as it does so, and my computer cannot turn down my central heating when the room reaches the required temperature, nor can it switch from one type of stitch to another.

Thus, it is possible to conclude that the machines I have spoken of are only capable of taking in certain types of information, processing that information and issuing the

output; they are, as already mentioned, machines with a finite number of possible states. They are to a very limited extent aware of their environment, (for there are only very specific things that they are designed to react to), and behaving as-though they understand the information they take in and process. The illusion that they have the ability to understand is promoted by the consistency of their processed informational output that is, by necessity of design, in keeping with the original informational input. Their behaviour may give the illusion of being complex and the product of a machine that must have a complex architecture, but it is only the product of a machine that has been programmed to be sensitive to specific informational cues that are received from their restricted environments.

Another form of behavioural complexity is the sort of 'second-hand' complexity of the design behaviour of the programmer who is writing some software for a computer to run and a person to use. The programmer has to consider the architectural complexity of the computer and the limitation of its capabilities, he or she has also to be aware of what the user might and might not be capable. This means that not only is the programmer trying to express the complexity of his or her own creative thoughts, but also the complexity of the possible users, the interface with the users and the sort of computer in which the software is to be used.[4]

In non-human animals there is a huge range of possible behaviours. All animals have to be able to process information if they are going to be able to survive. Indeed, all animals need to be able to process information in a very short space of time, what computer scientists describe as *real time*, for the decisions they are making are truly life and death. Any animal that is not aware of the danger in its environment, or that reacts too slowly to that danger, will not have the chance to run away again. Some fundamental element of understanding must be present that enables the animal to make decisions between what is and what is not dangerous in its environment. Similarly, animals have to be able to distinguish between those things that are good to eat and those things that are poisonous, and it is certainly the case that animals very rarely

consume anything that could kill them. But in non-human animals this might be a matter of their genetics rather than an understanding of their environment.

When it comes to defending themselves some animals have evolved highly complex and colourful displays that frighten or mislead possible predators. When threatened cats' fur stands up and they bare their teeth when danger is anticipated so that they can look twice their size and much more fearsome. Other animals rear up to increase their size, and some become as small and inanimate as possible, such as hedgehogs, so that the predator might fail to notice them or leave them for dead. Many wasps and flies without any stinging ability still have brightly striped backs, like those of more ferocious insects, so that they can look to all the world like predator and not prey.

It would certainly be admitted that many animals behave as though they can do more than merely process information, for they seem to be capable of understanding things in their environment and even knowing when it is best to run away or best to stay quite still until the danger passes. It might well be argued that they seem to know of needs for their own safety and that these judgements will have to contain an idea of how they see themselves in relation to their world.

I think it would be difficult to deny that these capabilities are conscious or deliberate for there is an element of judgement in them, albeit a split second one in the decision to fight or take flight. However, it would be much harder to make a claim for self-consciousness in these capabilities. But there is certainly a sense in which the animal knows that it is 'it' that is in danger or 'it' that needs to be fed. So it might be argued that it is self-conscious but not in the linguistic sense where the animal would say to itself, "I know that it is me that is being chased". That the animal can behave intentionally is one thing but that it might also being able to describe its behaviour using propositional attitude statements is another thing altogether, and one that I would argue is extremely doubtful.

Hintikka has argued that there is nothing added by, for example, my knowing that I know that Y; but I would argue that there is, and it is a proof of self-consciousness. For when "I know that I know the name of the woman I have just passed" then I have

reason to suppose that I will shortly remember her name. It is a sort of psychological "knowing" that permits me to reaffirm things to myself. It is not a logical one in which there exists nothing more than mere reiteration. Of course, if I never do remember her name then it might be said that I had mistaken her for someone else, or that I had simply forgotten it; but in neither case is it a matter of an error in my logic.

I think it is not possible to extract the behaviour that is exhibited by a system from the environment that the system occupies. No behaviour is exhibited in a vacuum and so I think it can safely be concluded that all our actions are, if not a product of our interactions with our environment, then at least directly related to it. Examples of this can be seen in the cases of prisoners who are kept in solitary confinement so that they have only the barest of links to their environment. Such people have been known to turn inwards, living in their minds and imagining scenes in which they would like to participate. Sometimes the imagined adventures become indistinguishable from their actual life and they lose their grasp of reality. The capability of imagining possible worlds is something that can be done only when there has been one there once on which to base the imagined possibilities. The capability to imagine is a function of an original environment with the added constraints that have been placed on the individual by his or her present environment.

Thus it can be said that the sorts of capabilities possessed by a system are not just a reflection of the complexity of that system but of the complexity of the system plus its environment. With this in mind I would now like to look at the third category of complexity in which complexity is seen as a product of the interaction between the system and its environment. Indeed in all the examples I have discussed it is hard to see how any of them could behave in a way that would not somehow, even tacitly, include their environment.

4.2.3. Complexity as the product of the system and its environment

The environment has been present in all of the examples thus far discussed, what little extra there might be is the question of just how much the environment influences,

or is present in, the capabilities of any system. It is certainly the case that the capabilities of all systems are constrained by the environment in which they live, or in the case of inanimate systems, the environment in which they are situated. For a situated automaton, such as a thermostat, the fluctuations in its environment influence its actions to the same extent as its internal mechanism, for it has nothing else. If either part was disabled the thermostat would cease to function. A computer is situated, but it has a more varied environment which is not rendered useless if one part of its environment fails to function; for example, if the mouse button is disconnected from this computer I am still able to move the cursor arrow by using the cursor keys. Similarly, if I remove a drawing application I am still able to use any of the other installed applications. Each of these two situated machines behave in the way they do because of their link to a specific part of the environment. And although the first is more constrained than the second, in respect to their environment, their capabilities are completely linked by their architecture and their relation to that environment.

Being non-situated, or free to move around, allows for an ever changing environment and an enriched informational input. To cope with this the capabilities of the system have to be much greater. For all animals the environment is important and their physical form and capabilities have adapted and evolved to suit that environment. The finches discovered by Darwin on the Galapagos Islands are good examples of this type of continuing adaptation. The finches have evolved beaks in a variety of sizes and shapes so that each of the groups of the subspecies can feed off a different set of plants and seeds and thus the species as a whole can survive. So in this way their complex evolution and continued survival is a product of their environment and their capability to adapt.

The capabilities of many other animals have been altered by changes in their environment. With ever growing towns the countryside is fast disappearing and foxes, squirrels and badgers have all become adept at foraging for food in their new urban environment. They have had to learn a whole new set of signs for danger and for food. But although their capabilities have adapted it does not mean that they are any more

capable than they were before. What they and human beings possess is the capability to adapt to unpredictable changes in their environment, and it is this capability that machines lack. A situated system has a predictable environment and anything beyond the range of that environment is beyond the scope and 'awareness' of the machine. A non-situated system has an unpredictable environment all of which is within the possible perceptual scope of the system; it is all possible informational input for the living organism.

But differences exist between the adaptations made by species and those made by individuals. Individuals adapt to changes in their own, personal environment and changes that affect only them or a group of people in a similar predicament to them. For instance, a group of people who feel saddened or angered by the depletion of mineral resources and the destruction of the world's hard wood forests might choose to change their behaviour so that they no longer buy products made from that wood and recycle the minerals such as steel and aluminium that are used to make cans for food and drink. The adaptation of a whole species happens on a much grander scale such as that made by the finches mentioned above and they are not something over which the species, or any member of the species, has any direct conscious control.

Animals or human beings have to be continually aware of their environment and flexible enough to select the pieces of information that are the most relevant to them and attend to them. They have also to be capable of deciding whether to act on that piece of information or select another piece. All of this has to be done in a tiny space of time to enable the system to do what is best for its continued survival. In the case of non-human animals the decisions are conscious but perhaps not self-conscious. Human beings, on the other hand, are capable of awareness of their environment on a grand scale and on a personal scale. They are capable of selecting information and processing it. They can understand the information they receive and extract knowledge from it. From this knowledge they are capable of forming beliefs about their world and about the worlds of other people. They can then choose to act upon those beliefs or ignore

them, for many people have principles by which they try to live but in unsuitable circumstances such rules can be overlooked.

On top of all this human beings are capable of making themselves comprehensively understood, and of understanding others in their social environment by using a shared set of symbols, a language. As was seen in chapter three the development of natural language has been possible for two reasons, the first is that human beings are by nature social animals with a shared environment and the second is that they are sub-symbolic systems capable of creating abstract symbols and assigning meaning to those symbols to produce a symbol system or language. Because of this human beings are capable of describing and discussing every aspect of their interaction with their environment, from the kind of weather we are having to a personal belief in a superphysical deity. So human natural language is a product of the relation between what we know to be a complex system and its environment.

4.2.4. A summary of complexity

I would say that the capabilities of computers, although vast and on the increase, are dictated by a combination of their internal design, their architecture, the program that has been instantiated and the environment in which they are fixed. They have no flexibility to choose what information to react to in their environment. The capabilities of non-human animals are still widely dictated by their environment but the higher-order animals, at least, have the added capability of being able to choose what they attend to in their environment. From this selection the system can choose how it responds to the information, that is, whether it will run away, fight, or conceal itself. This response may be dictated entirely by the arrangement of its genes, but with animals such as monkeys and even cats, it is not at all easy to rule out the possibility of there being a self-conscious element in their judgements.

With human beings it is easier to say what is possible for, as a member of that class, I have my own experience to go on. I know that I am capable of processing vast amounts of information, selecting what are the most important pieces for me,

responding to them and storing what is not immediately required for later use. The biggest advantage I have is that I know I do all of this, or at least the parts that necessitate it, with myself at the heart of my judgements. It seems reasonable to extrapolate from my own experience to say that other human beings have the same, or at least roughly the same, capabilities. I interpret incoming information subjectively and thus everything I explain to other people will have my own personal slant or interpretation. I am a product of my environment, but my many and varied capabilities are the product of my being able to see myself in my environment and act, for at least some of the time, for my own best interests. However, as a self-conscious yet social animal I am also capable of subjugating my own interests to the interests of the continued survival of society as a whole.

Thus I am capable of a high-level awareness of my environment, the selection of information from that environment, understanding the information and making self-conscious judgements involving it. I am also able to anticipate how other objects and states of affairs in my environment will be affected by my judgements, and to change my judgements or try to justify them to others, or indeed myself, using language. I, and all other human beings, if my extrapolation from myself as an example of human sentience and experience is correct, are very complex systems indeed with a great many capabilities.

In chapter three I discussed the methodology behind mental attitude ascription, and through the discussion it was demonstrated that the notion of ascription is a very complex one for which a complex system with many capabilities is required. Such a system has to be able to both identify the signs that imply the occurrence of a mental state and use language, or another form of behaviour, to ascribe the state. I will now look at the aspects of complexity that arose in the context of chapter three.

4.3. Complexity in the ascription and possession of mental states

In chapter three I looked at why, when and how we ascribe mental states to other systems; or in the language of cognitive science, what are the conditions under which I,

the observer, can decide that something I am observing has a particular inner representation. The conclusion I reached was that because we can never become that other thing, or even look inside its head, I can only ascribe inner mental states to it by looking closely at its behaviour and examining it for appropriateness in a particular context. From its behaviour I should recognise signs that are consistent with other, analogous mental state behaviour that I have already established in myself as the only true mental system that I can be said to know. The evidence for mentality can never be one hundred percent sure, thus the outcome has to be that even if another system does have mental states only it can know for sure that it has. This outcome applies across the board to all systems, organic and inorganic.

This ascriptive procedure is, itself, a very complicated procedure that is made up of a number of discernible complex actions. I shall discuss each of these separate aspects of complexity in the next section.

4.3.1. Creation of a paradigm case

Firstly, when I looked for a behavioural paradigm case, with which to compare other behaviour to see if it was indicative of the occurrence of mental states, I turned to the human system, and in particular myself, that is the only accepted possessor of mentality that I can ever hope to know with certainty. It is the most complex system of which I have a comprehensive, but still by no means entire, knowledge. It is possible to recognise from both what we know and what we do not know of the human mind in general, that we are dealing with a system of an amazing complexity. A system that is capable of conforming to accepted patterns of behaviour in an effort to understand and be understood; but equally able at deception and behaviour intended to mislead.

Any comparison, and setting up this paradigm case is no exception, necessarily involves looking at the system within its environment; for it is only when I behave in a particular way and I see other systems behaving in a similar way, in a similar environment that I can compare the probability of the mental states underlying the behaviour being the same as well.

4.3.2. Consciousness and self-consciousness in the environment

Other aspects of complexity can range from being conscious to being self-conscious in an environment. Part of being self-conscious is the ability to adapt and survive within a continually changing world. It could be contended that a desk-top computer can adapt to changes in its environment, but I would argue that it has been programmed to do so and its environment is a limited, finite one in which it has a great many possible states but all of them fixed by its program and ultimately predictable. What seems paradoxical about the human, or indeed other organic systems, is that the more we discover about them the more complex they seem to become and the less we realise we know. This is not the case with a computer that processes things serially because everything that it can do has to be known ahead so that it can be programmed into it. It is a finite state machine whereas the human system has an infinite number of possible states.

4.3.3. Language use and self-consciousness

The ability of human beings to report incidents and information from their interactions with the rest of the world is one of the most complex actions they perform. One of the prerequisites of being able to use language is that the system is capable of seeing itself in relation to events in its world. The advantage that sophisticated language users, such as human beings, have is that their descriptions of events and states of affairs in the world have a subjective element that a purely functional description does not possess. The notion of subjectivity is not an easy one, but basically it can be explained as how the individual sees him or herself being affected by events, or even possible states of affairs that might some day hold, in the world. For instance, if I hear a Party Political Broadcast on behalf of the Conservative Party I might justifiably form the opinion that their policies would be detrimental to my continuation in academic work. The result of this will be a personal or subjective belief that I ought to vote in an effort to change the government to one under which I, and others with interests similar to mine, would be better off.

If I were not able to see things happening in relation to me they would either have no effect on me, effect me but I have no forethought, or my report of them would be a purely functional one. A functional report is the kind of message that comes up on a computer screen to say that there has been a system error, or that more memory is needed, and so on. It is not the sort of thing that is subjective or even indicative of any form of consciousness, it is a programmed reaction to a particular set of circumstances; and is fundamentally no different from, for example, switching on the lights for the opening of the Christmas Season at Harrods, or waiting for the traffic lights to change at a crossroads.

So, as we also saw in section 4.2.3., one of the most significant signs of complexity is the possession of a reflexive relationship that pertains between the system and its environment. However, actually pin-pointing what counts as reflexive behaviour is, as we have seen, a very difficult thing to do. Indeed the only signs we have to rely on are either behaviour that is in keeping with having a mental state, or a linguistic utterance that professes the occurrence of a state of mind. These can either describe the situation, or at the very least demonstrate an awareness and possible understanding of it.

Of the appropriate behaviour and the use of language the latter is a surer sign of a system's complexity for, if the language is used correctly, it is an indication that the system is capable of possessing a great diversity of mental states, from the simplest information processing to the self-conscious formation of attitudes and the complex state of holding beliefs. These are the very states that are the product of informative interaction of a human being within its social environment, and they are those that have made it possible for the system to form a set of subjective beliefs and ideas about his or her world. However, much more is possible, for with language not only am I able to create and adapt my own set of beliefs, but I am also able to modify, by linguistic means the intentionality of other linguistic systems. Rational argument is just one example of how this type of modification can take place.

The social relationships within which we acquire and practice our use of language are themselves very complex. In them we have to do many things and exercise many of our mental and physical capabilities. Initially we have to be systems that have a potential capability for learning language, and on top of that we have to be capable of being innovative with that language so that we can create new words and phrases to describe novel situations that arise and those that might arise given the right circumstances. We have, in effect, to be intelligent 'sub-symbolic systems' capable of forming and using language.[5]

That our language arose in the first place suggests that there was a need to express more and more about events taking place in the environment and our relationship to those events. One suggestion might be that the purely physical behaviour through which we initially conveyed information might have become outmoded with the development of our environment and progression of our society with which it coincided. With a greater self-awareness our methods for conveying information have had to become more advanced and the sounds we make have taken on forms that enable us to explain and describe complex states of affairs.

Occurrences similar to the evolution of language have arisen in the rest of the animal kingdom. All animals communicate with each other in some way, and some animals are even capable of communicating with human beings. Of this latter kind I am thinking mainly of domesticated animals such as cats, dogs, and even some farm animals that have a lot of contact with human beings. The distinction drawn here is between organic systems that are capable of intra-species communication and inter-species communication, respectively.

It is not clear whether or not domesticated animals treat human beings as a totally different species from themselves, or whether they simply consider us to be extensions of themselves that provide some of the resources they would otherwise have to supply for themselves. This is by no means an easy question to resolve, but from a first hand example of one of my cats asking me, but not another cat, for food it would seem it makes some distinction between what it can reasonably expect from me and what it can

expect from another cat. It may just be a behaviour it has learned through repetition, but just what the distinction is it is impossible to discover.

There are many examples of animals that communicate only within their own species. One of the best is that of bees that have an extraordinary range of dances with which they are able to exchange information. Indeed entomologists have so far identified as many as sixteen different very complicated bee dances; but, as yet they have only discovered interpretations that fit three of the dances. In general, the dances allow a worker bee, having discovered a good source of pollen, to describe the location of the pollen to other worker bees. It seems that the dances are rather sophisticated for they have implicit in them the details of the directions and distances of the pollen from the hive.

For our present purposes what is most interesting is that their social interaction with one another has become elaborate enough to demand the development of an enhanced communicative techniques. It is vast and complex in relation to the magnitude of the bees' environment. But in comparison to the language of the human species it is an extremely limited form of communication. This is because there are great differences in what each species demands of its communicative process.

That the human system requires a language is indicative of the great complexity of human society. That a language is created, (and there are a great many languages), used and developed over an extensive period of time establishes the human system as one of the most complex systems that can be imagined. Our language facilitates the expression of a great many things, from straightforward descriptions of physical objects and states of affairs to the most introspective and fraught of our emotions; still further it can cater for the discussion of abstract concepts and ideas of philosophy, mathematics, theoretical physics, and so on.

The means of communication that is developed by a species can be seen as a reflection of the complexity of the social, natural and mental or cognitive environment of that species. With its very complex mental and physical states the human species has

developed a descriptive language that is accordingly very complex. This then is a reflection of the total environment of the human system.

4.3.4. The apprehension of complexity by the human being

The notion that complexity becomes an issue when we consider the way the human being apprehends complexity in other systems or states of affairs in its environment is one that we shall return to now. It appeared first in the context of how and why we ascribe mental states to some things rather than others. In this chapter I shall discuss two aspects of the human apprehension of complexity. The first is the human ability to apprehend other entities in relation to itself whilst also seeing itself in relation to the world. So that there are two stages of recursion required in this reflexive relationship between the human system, the system with which the human being is interacting and the world. The second aspect to be considered is the complexity of the decision-making by which the human being is capable of comparing the behaviour of one system with a system that is already known to be more complex than the first and conclude that the the one being compared is, or is not, itself a system as complex as the one with which it is being compared.

Within the recursive interaction I am assuming that it is acceptable that the human system is conscious; what I would add to this is the ability of the human system to be *self-conscious* in their interactions with other systems that exhibit *awareness* and also conscious of the 'I' in relation with the wider context of our world. This is no different from what many people already attribute to other systems such as computers and cats. In the case of a computer, its 'awareness' of its environment can be demonstrated through its reaction when someone types a command-line or clicks the mouse button. Cats can be seen to be aware of their environment in many ways, their posture, the movements of their ears, one eye being half open to keep an eye on things when resting, and so on.

However, problems arise, firstly in what way is the awareness that a computer has of its informational input different from the self-consciousness that is characteristic of

the human system; and secondly, what makes the possession of self-consciousness an advantage. As already mentioned above and in chapter three the possession of self-consciousness as it occurs in the human system is identified in two ways; firstly, by the use of propositional attitude statements by the individual that place him or her in a direct first person relationship with their world and the states of affairs in that world; and secondly, in a much less reliable way, by the careful examination of a system's behaviour to identify a correspondence between its behaviour and its ascribed mental state. The 'correspondence' is a matter of the consistency and appropriateness of the behaviour.

Propositional attitude statements are usually made in relation to something of which we can have a mental "picture". For example when I say "I believe it is raining today", I have a "picture" in my mind of the falling rain. And if I say "I hope my cat will come home soon" then I again have a "picture" in my mind of the return of my cat. Being capable of intentionality means I am able to interpret information I receive from the world outside and I am also able to conceive of new things that I 'hope', 'doubt', or 'fear' might happen.

I am able to recognise machine 'awareness' of its environment by its acting in accordance with information it has been given or with how it has been programmed to react to changes. For instance, if something does not match with what the system is programmed to expect it will give an error message that tells me what is required before it can proceed. So the machine is also capable of acting reflexively within its environment. What then, if indeed anything, are the differences between machine reflexivity and human reflexivity?

If we look back to chapter two, and paragraph two of section 2.11. there is a brief statement about the distinction between organic systems and machines. The former have the capability for self-conscious introspection, whilst machines may be capable (depending on the complexity of their internal structure) of reflexive activity and the outcome of this might be some type of emergent property or properties. However, the

result of reflexive action within the machine is more likely to be a predictable behaviour or sequence of behaviours that are the result of the instantiation of a specific program.

Self-consciousness allows the human system to look at itself in relation to its environment, but strictly speaking this is no different from what a recursive program can do in a machine; but it would seem that there must be something more going on in human self-consciousness for intuitively I can tell that my awareness of my environment is richer and more diverse than that of the program in the machine. We set about recognising self-consciousness and recursion using the same techniques with which we learn to recognise and identify mental states and intentionality in systems other than ourselves; that is, by the examination of behaviour and the corroborative use of propositional attitude statements.

Here the differences begin to show for the machine is constrained by the limitation of its scope or capacity for experience. On top of this there is the fact that it runs on a basis of programmed, and not naturally perceived and processed, information. Thus even if a machine, using a voice synthesizer, could utter propositional attitude statements they would still have to be written into its program for such utterances cannot be made at whim by a machine.

Another way of thinking about this notion of choices that are made 'at whim' is to consider the notion of *subjectivity*. A machine cannot subjectively choose to do something that is not already part of its physical structure or dictated to it through its instantiated program. Within certain, perhaps physical, limits the human being can choose to behave as it pleases. A human being can choose to be moody and unpredictable even though his or her life is successful and all the outward signs would suggest that they should be happy. In an individual's choices there is an element of subjectivity that allows his or her personality to be expressed.

The choices that I make are influenced by the experiences I have had and no other person can make the same choices for no other person can have had all my experiences, and historically, physically and mentally there is only one me. Many machines can have the same structure and internal design so that any two machines with the same physical

structure and the same program will always react in the same way to the same problem. There is no subjective element at play here, for the machine cannot say to itself "What would 'I' like to do in this instance?". There is no concept of 'I' in the course of action that the machine follows, whereas because I am self-conscious I see myself in all my judgements. I am able to see myself in relation to how I remember the past, how I deal with the present and how I foresee my future. I am conscious of the very complex relation in which I stand to my world through time.

It could be said that the machine is aware of its incoming informational input or stimuli in much the same way that I would be aware of an electrical shock or impulse that is passed over my skin. I react to the impulse. I do not respond to it for implicit in the notion of response there is the suggestion of something more premeditated and thoughtful. When I feel the sudden twinge of pain I withdraw my hand in an impulsive or innately dictated action. I automatically flinch from the pain without having any choice in the matter. I do not have to weigh up whether or not I prefer to withdraw my hand or leave it there to sustain further injury. In just this manner the machine reacts to the relevant incoming stimuli, it does not sit and muse about the outcome of its reaction, it simply reacts. The machine might be aware of the sensation or stimulation but it cannot feel the pain in the way a human being or an animal can. However, neither of these ideas for a distinction is particularly novel; they can be found in Stanley Rosenschein's recent work.

The human abilities to be self-conscious, behave intentionally and ascribe meaning to symbols are linked in at least two ways that are the inverse of the machine constraints already mentioned. The first is that the human being is not a passive receiver of information, and the second is that it cannot help but ascribe meaning to the events and states of affairs that it encounters. In the first case human beings actively go out seeking information and the information they find is always processed subjectively with the 'I' of their self-conscious judgement always being present. The second case is slightly more awkward for it requires that the human system be linguistically oriented.[6]

Selectivity and flexibility

As an organic system at the top of the phylogenetic scale the human being is capable of moving around its world in search of all kinds of information. It is therefore capable of actively experiencing a continually changing set of events and states of affairs with which it has not previously been acquainted. It will have no use for a lot of the new events it experiences and these are either banished to the realms of peripheral perception or ignored altogether. The information that is important to the individual is processed through the senses and either used or stored for use in the future. But the fact that human beings are not passive receivers of information means that they have the ability to choose what is important to them and select only certain pieces of experience or information for special attention.

So what we have so far is that the mental life of human beings is in a continual state of flux and the human system is a very complex one that can deal with there always being new and different stimuli to attract its attention.[7] To say that it has the capability to select those pieces of information which are of use to it whilst ignoring others is to suggest that the system is very flexible in its approach to the range of incoming information. When an initial selection has been made the range of information will have been narrowed down and from this it is possible to select specific objects or events to which the system can give yet closer attention. It is these selected pieces of information Dretske describes as 'digitalised'.[8]

Such pieces of information are selected by the human being on the basis of what is most appropriate for its well-being; but they can also be for its amusement. One example of this would be that I can choose to disregard a sensation of hunger if I am enjoying a conversation with friends or engrossed in reading a book. The scope of choice that a human being has is immense and its level of flexibility to select the most appropriate information has to be correspondingly great.

Selectivity is based upon the flexibility of the system that is doing the perceiving and the breadth of the range of its possible choices. A machine can be said to have a

limited scope because it is only capable of following a course of action that has previously been sketched out for it in its internal design and program. It cannot exercise any capability to choose. Any system that has alternatives, although those alternatives might be limited, and is capable of choosing the most apt of these in a given situation, is exercising a greater amount of flexibility of choice than a computer with a fixed program and structure.

For example, a thermostat has no flexibility and only a very limited scope for the receipt and processing of information. It cannot demand food, it cannot feel tired, and it cannot converse about its perceptions because it has no use of language and it is not aware of anything other than that which it has been programmed to perceive.

In between the two 'extremes' that I have chosen, of human beings and thermostats, there are a great many other different systems, both organic and inorganic. A machine with more capabilities than those of a thermostat would have more incoming stimuli and it follows that the system itself would have a greater flexibility, however its choice of information is still dictated by the program it is running at the time. On the other hand a cat can choose between all sorts of incoming stimuli in its environment and unlike the thermostat it has no fixed or 'situated' environment. It can move around its world, in much the same way as the human being, seeking new information about food, territory and possible mates. It has a greater flexibility to select the information that is of immediate importance and thus narrow down the field of relevant information.

Assignment of meaning

Embodied in the notion of being flexible in the selection of the most relevant incoming information is the notion that human beings ascribe meanings to events and experiences even though they may not be consciously doing so. For instance, when I look at clouds I often interpret their form to fit something with which I am familiar, often seeing human faces or the shapes of animals in them. The same thing happens with doodles or scribbles.

In *Hamlet*[9] we can see an example of just this type of occurrence. In the dialogue between Hamlet and Polonius that follows it is possible to see the many forms that Hamlet, whether in real or feigned madness, imagines a cloud can take.

> Hamlet: Do you see yonder cloud that's almost in shape of a camel?
>
> Polonius: By the mass, and 'tis like a camel, indeed.
>
> Hamlet: Methinks it is like a weasel.
>
> Polonius: It is backed like a weasel.
>
> Hamlet: Or like a whale?
>
> Polonius: Very like a whale.

We can also see that Polonius tries to look for similarities between the forms that Hamlet is reportedly seeing and the clouds that he can see; and it is nearly possible to see Polonius convincing himself that "Yes, he's right, if I look at it this way that cloud is very like a weasel".

A similar thing happens when we read tea-leaves in the bottom of a cup or we have someone tell us our fortune from the laying out of a set of cards. In each case a meaning is attributed by the reader and another level of meaning is attributed by the person for whom the fortune is being read. This new level of meaning is constructed by the person whose fortune is being read their own knowledge of their personal history and from this extra information the fortune-teller's interpretation can now take on a new and enhanced meaning.

Seeking examples for the attribution of meaning to events and states of affairs in our worlds is by no means difficult. Indeed, for suitable examples we need look no further than mythology and the theory of *animism*. In the former we see natural events, such as thunder and lightning interpreted as, for example, the wrath of the Gods to instil fear into the hearts of mankind. And, in *animism* we find explanations in the form of the ascription of intentionality by young children who relate events in the world to what is meaningful to them. So the sun rising and setting becomes "The sun is getting up" and "The sun is going to bed". Nothing remains without meaning for too long in the human world because, in much the same way as Dennett describes the *intentional*

stance as a predictive tool, we make sense of everything we encounter so that our interactions with objects and events that are external to us are made easier.

It does not matter that our ascription is not completely accurate, for what is important is the reliability of the prediction. For the child the sun's 'getting up' and then in the evening its 'going to bed' is a useful way of establishing a continuum of complex solar events whilst not having to understand any of the difficult scientific concepts of space, time and motion. Quite simply the child compares the movements of the sun to her own daily events thus enabling her to predict a future state of affairs. Their explanation is meaningful to them, and it works even though it is not scientifically accurate.

I will now summarise what has taken place in this section and then move on to look at the promised second aspect of the complex notion of how we, as human beings, decide that another system is, or is not, itself a complex system.

An interim summary

This section dealt with the difference between the complex self-conscious relation to the world and a simpler relation of awareness of the world. The observation of appropriate interactive behaviour between a system and its world is the only way we can infer its awareness. For a human being this behaviour can be purely physical interaction or linguistic interaction in the form of propositional attitude statements - both being forms of behaviour. For a machine such as a computer the behaviour is the reaction between what is typed on the keyboard and what the computer proceeds to do in accordance with the command it has been given.

Human beings can choose how they wish to act and no two choices made by any two human beings will ever be exactly the same for they can never have wholly identical contributing experiences. A computer does not have any subjectivity or element of choice in its actions; which is to say that all its actions are the result of internal design and instantiated program.

In the case of higher-order animals the lack of evidence makes it difficult to know whether their behaviour is a result of self-conscious choice or just innate drives. In *The Emperor's New Mind* Penrose appears to come down on the side of self-conscious higher-order mammals, such as monkeys. The example he gives is of a monkey that is trying to reach a banana that is hanging from the ceiling. In the room with the monkey is a box and after some fruitless (sic) attempts the monkey displays a sense of realisation and brings the box over to just below the banana, climbs the box and takes hold of the banana.[10] If we take into account the close relationship of humans and monkeys and we accept that the criteria for recognising and identifying behaviour in humans can hold for monkeys, then it does appear from this example that monkeys, too, are self-conscious and have the ability to make subjective decisions.

Implicit in having self-consciousness is the notion that the human being is able to see him or herself in relation to their world, and with such a capability come the attributes of being flexible enough to select from a huge range of possible choices the course of action that will best suit the individual in his or her bid to survive. The flexibility to choose diminishes when we move to simpler systems, indeed when we reach a system such as a thermostat the choices are non-existent. Systems that are not fixed, such as robots, cats, and monkeys have to have greater flexibility to select things to attend to in their environment. The robot, of course, is still limited in the sense that it can still only do those things for which it has been designed, but with it being capable of movement it will have more inbuilt decision making mechanisms.

In relation to these points if a system other than a human being were self-conscious or flexible enough to select any piece of information for attention, it would not be able to tell us for, as yet, only the human being is capable of assigning meaning to symbols and creating language in just the way that we have. In fact assigning meaning is not only confined to straightforward symbols, for human beings assign meaning to everything in an effort to understand and predict their world.

Thus I would conclude that the relative complexity of a system depends upon (i) the amount of information it processes, (ii) its flexibility to select the most relevant piece of

information, (iii) whether the choice is made on a basis of a set of programmed instructions or subjectively, (iv) the system being able to ascribe meaning to incoming information, and finally (v) being able to use language to describe its relation to its own world. A very complex system is one that can fulfil all of these criteria and the only one that can, as yet, do this is the human being.

The complexity of the decision making process

In chapter three I spoke of the ascription of mental states being dependent upon how complex we think the other system to be. The sort of apprehension we have of the other system depends upon the consistency[11] and appropriateness of its behaviour in a variety of circumstances. So there are three things that are required; firstly, that the human being doing the comparing has devised a set of reliable criteria upon which she can base her judgements, secondly that a justifiable link must exist between different degrees of complexity and the achievement of different levels of mental states, and finally that the human system is complex enough to be able to make a rational comparison between the established criteria and the exhibited behaviour.

The criteria for mental state ascription are set up by analogy with a paradigm case, and the best model to use is that of a system we already accept as possessing mentality of a sufficiently high, or even more complex, level than any other known system. In this instance then the most appropriate model is the human being. The analogy is not necessarily made using language, for just behaving in a particular way with something can show that we assume it has a certain set of capabilities. So the apprehension of at least some kind of mentality - or assumed mental state - did not start with language; however, the assumed mental state of another system is most accurately, although not necessarily, expressed using language.

By dint of their having no physical location in space and time, and therefore a somewhat idiosyncratic physical manifestation, the description of mental states is a difficult procedure. What we tend to go on are the "family resemblances" that are the most commonly observed features accompanying the possession of a particular mental

state. Of course, the advantage that the human being has is that it is capable of expressing aspects of its mentality, thus enabling the observer to compare the physical manifestation of a mental state with the verbal corroboration. In this complicated manner we established a set of analogous mental state behaviours. It is frequently a matter of trial and error when we try to understand the mental states of another system. Often we can be misled or deliberately deceived by behaviour that does not fit the accompanying propositional attitude statement, but by and large we get by using the analogies we have learnt through social interaction and the application of the context to the actions, for example, an actor who 'dies' in a piece of street theatre may move our hearts but we do not believe he is dead.

The second requirement is that there is a relation between complexity and the levels of mentality that a system can reach. I shall keep this discussion brief because I intend to say a lot more about this in chapter five when I take a closer look at two hierarchical arrangements.[12] A prerequisite of being able to show any relationship between these two things is that mental states can be divided up into levels of difficulty. So that we can say, for example, that being able to take in information is of a lower order than being capable of processing it, and being able to select the relevant piece of information from all sorts of stimuli is more sophisticated than having a limited environment and no capacity for selectivity.

It would appear superficially that this is not such a difficult thing to show, for in the example earlier of a thermostat which we know to have a very simple structure it was shown that its capabilities are limited, it has no other function than to process specific pieces of information and it has no freedom to exercise selectivity for it has no choices. A simpler example is an automatic kettle for it can switch itself off when the water reaches boiling point, but it cannot switch itself on when the water temperature drops below a hundred degrees Fahrenheit. So the kettle must have an even less complex mechanical structure than a thermostat.

Keeping to the same, previously encountered, examples, a video recorder has a greater capacity for informational input, and although it is still programmed the system

has a wider range of information from which it selects the pieces that take priority. In this way it is able to record programmes in the order in which they are screened and not get the dates mixed up and thus overlook or forget some.[13] Indeed if someone had a video recorder with the capacity to forget information it would be sent back to the manufacturers because it is unreliable! It seems then that reliability will be a trade off with the increase of complexity.

Looking at human beings and the amount of information they can take in, select, process, interpret, form beliefs about and explain using language, it is not difficult to see reasons for which we should accept them to be very complex systems with a great many functional capabilities. It is also possible that the diminishment of reliability is in keeping with the notion of human complexity, for so often we misunderstand information, mis-remember information or just plain forget it. We talk of *selective retention* being a feature of the human memory, what we mean is that human beings tend to remember things that have a particular relevancy for them and discard things that are irrelevant. But the process of forgetting is not always this methodical and useful things get misplaced.

By now some things have been set down as things the possession of which indicates a high degree of complexity. They are: self-consciousness which is best demonstrated through the use of propositional attitudes and thus requires the use of language; being capable of the selection of relevant pieces of information and the subjective interpretation of the selection; being able to assign meaning to symbols and to use language; and, finally, being able to forget information that is no longer relevant - a sort of selective process in reverse.

The notion of rationality is not something that need be confined entirely to the mental life of human beings for it would seem that there is nothing in 'being rational' that does not also exist in 'being logical' and, by their very nature, computational machines are logically bound. Even a thermostat functions on the basis of a binary code. In being logical or rational there is an element of being correct, even sober, in one's judgements; (though it must be remembered that 'sobriety' of judgement is a

noun used only when describing the decisions that human beings make because they are capable of rashness and insobriety). To be entirely rational in all one's judgements would mean the adoption of a complete impartiality that would sometimes mean that the end result is detrimental to oneself.

One's own subjectivity should not enter into a rational judgement for the decisions are both factual, that is, based on fact, and matter of fact. It is not so much that the 'I' is not present in the judgement, it is more that the worry about "how will 'I' be affected" is removed altogether. In this way then rational action is something of which both human beings and machines are capable. However, it would be mistaken to say they are equally capable for one is drawn to conclude that ultimately the machine is always absolutely rational since there is no possibility that its judgement could ever be clouded, even unwittingly, by introspective thoughts.

However, in the case of mental state ascription, being able to rationally compare two things in the environment is a different matter for it demands that the system is capable of seeing beyond itself and into the world where a comparison can be made and that the two things to be compared have some essential correspondence. This, in turn, requires an understanding of the things being compared; which in this case are appropriate mental state or 'human-like' behaviours and the possibility of concurrent mental states.

As already discussed there is no restriction on what the human being can perceive within its environment so it is free to compare the attributes and existence of any physical objects or states of affairs that it encounters. But, more than this, it is able to entertain the fundamental ideas behind abstract concepts and compare the outcomes of possible future states. Any machine, no matter what level of sophistication it might now have reached, is always constrained by its design which dictates those things in its environment to which it can react. Nor is a machine likely to be interested in what the 'good' life might be, or in whether or not any other system has mental states. It does not have to try to predict or contrive the best way it should interact with me.

Thus it appears that the creation of an analogy with my own mental states and the comparison I make between myself and the behaviours of other systems is something that is peculiar to me and human beings in general.[14] Only I wonder about the nature of other things in my world and only I, and other human beings, try to create analogies between our behaviours, experiences and possible mental attributes. It is not something that computers have been designed to do, nor is it something for which non-human animals have any obvious need to do since their interaction with the world is on a much more basic level.

This and all the other areas of complexity that arose implicitly in chapter three can be found as aspects of any one of three categories of complexity that were discussed at the beginning of this chapter. I shall now draw this chapter to its conclusion with a summary of the main points that have arisen and a look at what is in store in the next chapter.

4.4. Conclusion

In answer to the original question "given a specific task or competence, what is the minimum system that would be required to accomplish it", it can now be answered, if still only provisionally, that for a system to be capable of, for example, processing information it must first be aware of its environment. Such 'awareness' is demonstrated, even by the most limited systems, by their capability to react to stimuli that are relevant to it. However, being able to respond to a fixed type of stimulus, such as a rise in temperature, does not indicate that the system has any flexibility to decide which are the stimuli that are relevant to it. Indeed it suggests that the system has a very limited range of actions or behaviours and no flexibility at all. So a simple awareness only shows that the system can respond to the aspects of its limited environment for which it has been programmed.

For us to say of a system that it "knows X" the system would have to demonstrate first that it had understood "X". To do this it would have to explain its understanding and answer questions on its claimed knowledge. For example, when someone gives

you directions to get to the local library, you repeat their directions to show them that you have understood and that you now know the correct route to take. But this is an easy example for it is spoken verification. Other examples are purely behavioural and for these we can never be completely sure that any system other than ourselves does truly understand or know anything. Someone may pretend to understand and be lucky enough to nod in all the right places and so fool us into thinking they know what we mean. Machines on the other hand act in accordance with our requests and their programming and in so doing can fool a great many of us into thinking that they actually do understand what we are typing in. Maybe our criteria for what count as understanding and knowing behaviour are not yet precise enough and this is why we can be so easily fooled.

To be capable of making subjective judgements the system needs to be self-conscious, and being self-conscious requires that the system has the flexibility to choose those stimuli in its environment that are the most relevant to it and its continued livelihood. Making these sorts of judgement necessarily includes an element of subjectivity for each judgement will be made on an individual basis to fit a specific set of personal circumstances. The only system that we know for sure to possess such self-consciousness is the human system. Our certainty is based only on its capacity to report its intentionality and intentional actions using propositional attitude statements. We do not, and perhaps cannot ever, know whether or not animals behave intentionally. Behaving intentionally would permit the presence of self-consciousness in their actions and they cannot use language to inform us of its presence. Nor is it possible for us to imagine what it is like to be another animal in the way that it is possible for us to imagine what it is like to be another human being.

In the next chapter I will offer a further examination of the relationship between the complexity and capabilities of different systems. I will begin by looking at two hierarchical arrangements that have already been constructed. The first was developed by Chomsky to show an incremental relation between the complexity of a system and its capability to recognise and interpret different levels of grammar. In the discussion of

this hierarchy there will be a closer look at the actual physical requirements of four different types of machines that display different capabilities. The second hierarchy is one designed by Dretske that first shows a division of intentionality into three levels, and then goes on to demonstrate how each level relates to the complexity of a particular system and the ability of that system to process incoming information and possibly act upon it.

I will offer arguments to show that Chomsky's hierarchy is successful for he deals with machine states that can be quantified and a direct relationship can be shown between the machine and its capabilities with little or no difficulty. Whereas because Dretske deals with mental states his hierarchy is bound to fail for it is not always the most complex system that can carry out the most complex tasks and some very simple systems can do very complicated things.

Endnotes:

[1] 'Complexity' in this sense bears no relation to the technical sense of 'complexity' in 'Complexity Theory'. I mean 'complex' as in 'intricate' or 'not simple'.

[2] Penrose, R (1989) *The Emperor's New Mind - Concerning Computers, Minds and The Laws of Physics*, Vintage Press, p.483

[3] Ibid. p.551

[4] If design behaviour is second-hand complexity, then so too are thoughts or any explicit representation of objects or goals for the system. In this sense first-hand complexity is the physical structure of the system and its potential in relation to its environment; and any product of this is second-hand.

[5] Clark, A. and Karmiloff-Smith, A. (1990) The Cognizer's Innards,

[6] See above section 4.2.3., fifth paragraph.

[7] This is not to deny that other organic systems are complex; nor is it to deny that they have a continually changing set of stimuli in their environment. The comparison being set up is between the flexibility of the organic system to detect and select incoming information, as opposed to the inorganic system that depends upon its program and its fixed environment for information from which it has no freedom to choose.

[8] Also see Dretske referred to in chapter 2 and again, in greater detail, in chapter 5 section 5.3.

[9] Shakespeare, W. (1604) *Hamlet*, Act 3 Scene 2, lines 383 - 389, The New Penguin, 1980.

[10] Penrose, R (1989) *The Emperor's New Mind*, Vintage Press, p.550-552

[11] The notion of the 'consistency' of behaviour is really only useful when we are dealing with organic systems that possess a brain, for with known mentality their behaviour may be subjective and not dictated in the way the actions of a computer are dictated to it by the program it is running.

[12] One set up by Chomsky that shows a direct relationship between the recognition and processing of different levels of grammar by machines that vary in complexity, and the other is Dretske's relationship between levels of intentionality and corresponding mental states.

[13] The notion of forgetting is an interesting one for it suggests that being capable of forgetting takes a very complex system indeed. A system that has so much informational input that it forgets to deal with important things or gets its priorities wrong and fails to do things in the most successful order is a system that has a capacity for interesting behaviour.

[14] I think it is probably very unlikely that animals create anything similar to the rational comparisons that human beings make and this is quite simply because their lives do not demand interaction, with each other and with other species, on this sort of complex level.

5. A hierarchy of complexity and capabilities

5.1. Introduction

In this chapter I will extend the conclusion I arrived at in chapter four: that any system's capabilities depend upon the environment that it inhabits, and to show that a positive relationship exists between the complexity of a system and the complexity of the tasks it is able to carry out, (so that a system's capabilities are a function or product of its complexity plus its environment). By the end of this chapter I intend to have shown that hierarchical structures, relating the enormous diversity of capabilities of a system to both its architecture and its ability to adapt and generate behaviour, are sufficient for describing machine states but not for the description of mental states.

The ability of any system to generate behaviour can be divided into two types. The first of these is spatial in nature and remains statically present in the design of the system. It is not flexible and cannot be enhanced or altered in any way by changes in the system or the system's environment. It has no scope and is capable of generating only a limited set of behaviours within the system. An ideal example of a system that has the capacity for this sort of fixed behaviour is a kettle because no amount of change in its environment will create any change in its actions. The second type of generative capability is temporal in nature and dynamically generated in the system. It is flexible enough to be able to change itself and bring about changes in other systems, and there is no limitation set on this behaviour because the full structure and capabilities of the system are continually changing and cannot ever be fully known. Simple systems possess only the former whilst more complex systems have the capacity to possess the temporal generative ability since they are systems that need to adapt to survive within a continually changing environment.

This chapter will be structured in the following way. The form of previous chapters has been to begin by giving a statement of the problem and the area that will be covered in my attempt to deal with it, this I shall do in section 5.1.1. Then the discussion will be opened in 5.2. by setting out Chomsky's hierarchy. I shall first say why Chomsky's work is pertinent to my own work and then take a closer look at the grammars and the machines that Chomsky describes and the relationship that he claims exists between the two. Then I will follow the same procedure with an examination of Dretske's hierarchy of intentionality. Dretske's hierarchy is more obviously useful because he deals with intentionality and its relation to the possession, by a number of different systems, of particular mental states. I disagree with some of the points that Dretske makes and I will offer arguments to show why I do so and then go on to offer possible solutions to these difficulties. The chapter will be brought to a close with a discussion about why hierarchical stratifications are a successful way of dealing with machine states that are distinguishable and quantifiable but not of examining mental states which are vague and thus difficult to define. A new strategy for comparing machine states and mental states will be examined in chapter six.

5.1.1. A statement of the problem area

Initially the problem to be dealt with in this chapter is whether or not it is possible to demonstrate using a hierarchical structure that there is a positive correlation between the architectural complexity of a system, its environment and the functionality it possesses. By 'functionality' I will mean the things of which the system is capable. In chapters three and four I made use of many examples that showed that in our shared world there are a vast array of systems, both organic and inorganic, that are capable of behaving in ways that vary quite considerably; some are only capable of processing information, whilst others are capable of a full understanding of their environment and forming beliefs that will dictate and direct their subsequent behaviour. I will argue that the

complexity of the tasks or behaviour that a system can carry out will be directly related to the architectural and environmental complexity of that system.

I will turn now to an example of an architectural structure that was developed by Noam Chomsky in 1959. His architecture demonstrates the idea that the tasks that a system carries out are related to the internal complexity of that system and the amount of information that it has been designed to respond to within its environment, so that where a system has a simple internal design and a very limited environment it will be capable of carrying out only the simplest tasks. With an increase in the complexity of its architecture it is probable that there will be a corresponding increase in the number and variety of the stimuli within its environment to which it can respond; which is just another way of saying that it will have a greater functionality, or range of capabilities.

5.2. The Chomsky Hierarchy (1959)[1]

Still looming large is the question of when it is justifiable to ascribe mental states to non-human systems and in chapter three I argued that it is possible for me to know my own mental states and by my interaction with other human beings I can extrapolate from my experiences and their commensurate behaviours that they too have mental states that are very probably qualitatively similar to mine. However, when it comes to ascribing mental states to other systems things become a lot more problematic. It is no longer possible for us to reasonably say that we 'know' what it is like to be a cat or moose or any other animal, nor, for that matter, is it possible for us to know what it is like to be a machine such as a television or a thermostat. Our decisions about whether or not another system, organic or inorganic, has mental states are based upon two things; its actions and our apprehension of its architectural complexity. A positive decision about its possessing human-like mental states will usually depend upon how consistently human-like its behaviour is and whether or not we consider it to have an internal complexity that makes it possible for it to occupy a mental state. This 'human-

like' behaviour I have called 'as-though' behaviour, because the system behaves 'as-though' it knows, understands, and so on.

In the work that follows I will examine how Chomsky shows that there is a correlation between the complexity of a machine and the complexity of the grammar that it can recognise and interpret. He deals only with the capabilities of some machines and not with the capacities of any organic system so the field is already suitably narrow making it a good place to start our examination of hierarchical relationships because there are already boundaries or constraints placed on what we are to look for. We shall be examining the progressive changes in the capabilities of inorganic systems as their architectures become more complex and their connection to the environment becomes more enriched. So at this stage there is no need to go beyond these boundaries to look for implications about organic systems as well.

5.2.1. The grammars

In particular Chomsky wants to look at the ability to recognise and interpret phrase structure grammars, of different levels of sophistication, that are demonstrated by four different machines. What he offers is a straight-forward comparison, showing an incremental increase in the complexity of a system, the elements in its environment to which it is designed to respond and the capabilities that these two can together afford the system as a whole.

I shall make one or two general points about the grammars and then take a look at the four grammars and what they comprise in. Then I will move on to examine the machines that Chomsky uses to set up the other side of his hierarchical comparison. But before I do any of this I will offer a clear view of the hierarchy in the form of a diagram so that reference can be made to it as the text is read.

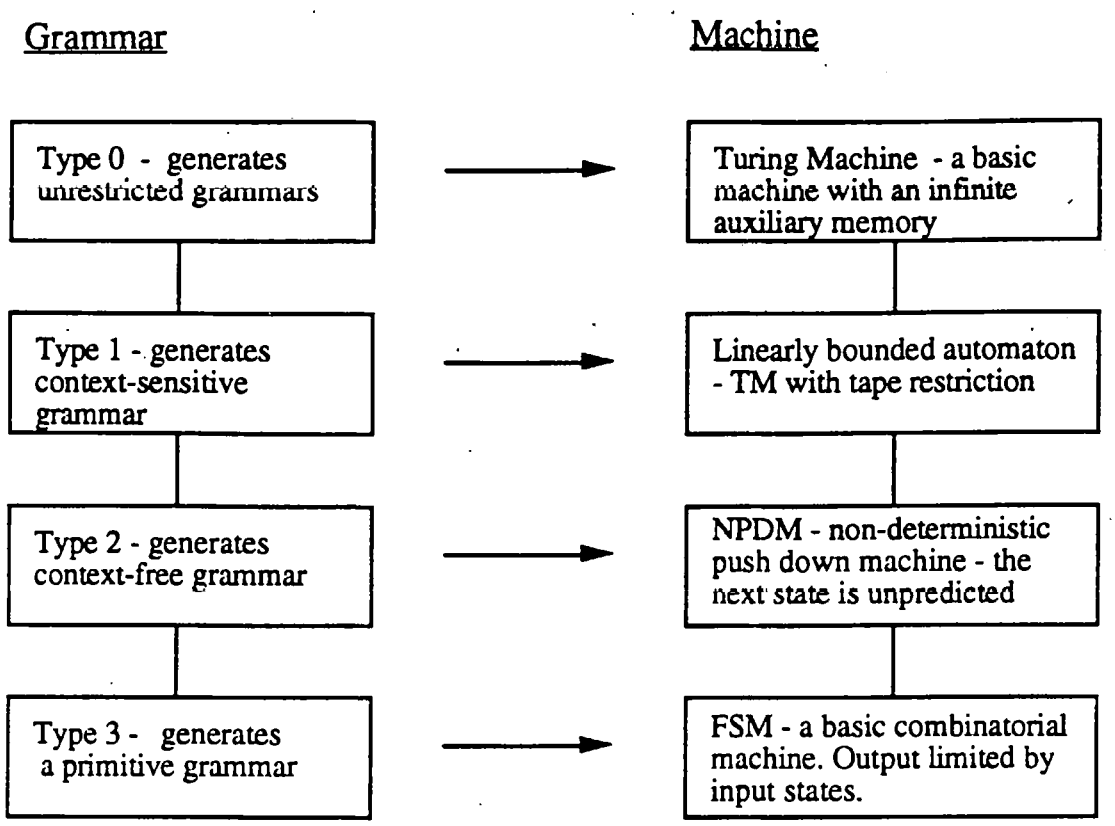

Figure 4

The sentences or phrases that a phrase structure grammar generates are called the *surface structures* of the grammar. In formal languages, such as those used for programming, the description stops at the surface structure. In natural languages descriptions can often go below the surface structure to *deep structures*. We talk of sentences that are ambiguous as having hidden meanings and in grammatical language these are said to have one or more deep structures that are below the surface structure. The constraints that are talked of are those that are placed on the production rules of the grammar. They produce restrictions and consequently make the grammar easier to understand.

The first grammar that we will look at, but the one that comes at the bottom of the hierarchy, is *Type 3* grammar, or the set of regular grammars. These are also known as finite state languages, where the finite states are equal to a finite set of nodes on a

transition graph.[2] Thus when we find ourselves in the middle of a sentence using this grammar the only information we need to know to enable us to finish the sentence correctly is what state we are in at present. No other information of any kind is necessary. So, for example, we do not need to know the content of the first part of the sentence that has already been written or the context of the greater piece of writing of which this one sentence is only a very small part.

An example would be as follows:

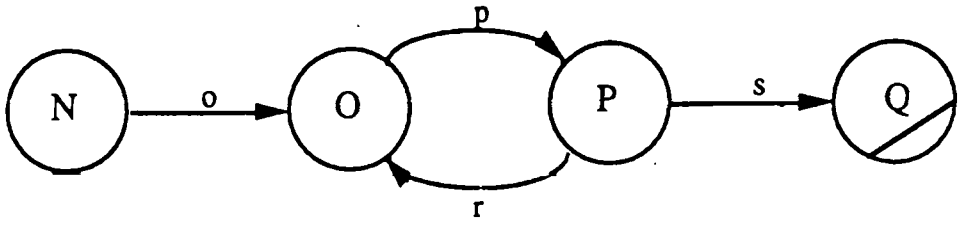

Figure 5

N, O, or *P* are all non-terminal symbols; the start symbol is *N* and is commonly the 'determiner', examples of which are 'the', 'an' and 'a'; *o, p, r* and *s* are terminal symbols and the final node is indicated by a diagonal bar across the circle. If *N* is 'The', *O* is 'ginger', *P* is 'cat' and *Q* is 'sleeps', then the sentences that could be generated will be sentences such as, 'The ginger cat' or 'The ginger cat sleeps'.

Such grammars and finite state networks as these offer a simple mechanism for the generation of sentences and the analysis of language. However, because the mechanism is simple it means that many more interesting sentences, and indeed, languages cannot be generated. An example that Krishnamurthy offers[3] is that finite state grammar finds itself incapable of dealing with sentences in English because of the richness and variety of the expressions that it uses. It finds parenthesized expressions difficult to describe for the same reasons, that is, that their meaning often depends upon the expressions within which they are embedded or they are simply asides about which the present state is unknown.

The regular expression language can be produced with a very basic finite state machine. It is used for representing regular sets. Three benefits are offered to the user of a language as basic as this. Firstly, its expressions can be written out in a line from left to right and it is then obvious which are the terminals and non-terminals, and which are the start and finish nodes. Secondly, it has a precision and formality that natural language has not. And thirdly, it is the most simple of all the formal languages for a designer to use. However, the third benefit has an accompanying drawback, and it is this; because it is the simplest language that can be used there are relatively few things of any great interest or significance that can be done with it.

The next grammar that Chomsky makes use of is *Type 2* or context-free grammar. These are used extensively to describe both formal and natural languages. These grammars are of a slightly more complex form because they are context-free and not limited in the way that a finite state grammar is. An example of their form would be, "A->x" where 'A' can be replaced by 'x' anywhere it appears for there is no constraint on the context. It is still a fairly simple grammar that is often easier to use than the more complex 'context-sensitive' grammars which are the next up on Chomsky's hierarchical scale.

By looking at the derivation of a sentence in context-free grammar it is possible to show how a particular sentence can be generated from its rules alone. The clearest way of seeing how a derivation of this sort operates is by looking at the diagram of a *parsing tree.*

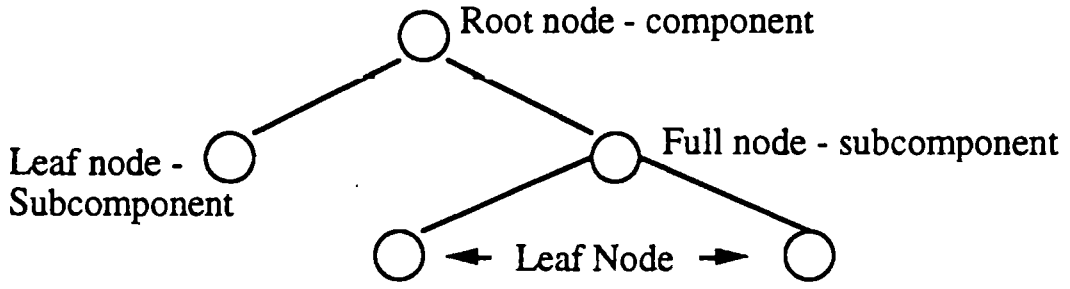

Figure 6

The start symbol always appears as the root of the tree, (which paradoxically is at the top of the diagram), and the terminal symbols are at the end of the branches just as the leaves are on a tree.

In the following diagram I use a sentence of natural language[4], "The man hit the ball", to demonstrate this more fully where 'D' is the determiner, 'NP' the noun phrase, 'VP' the verb phrase, 'N' the noun and 'V' the verb.

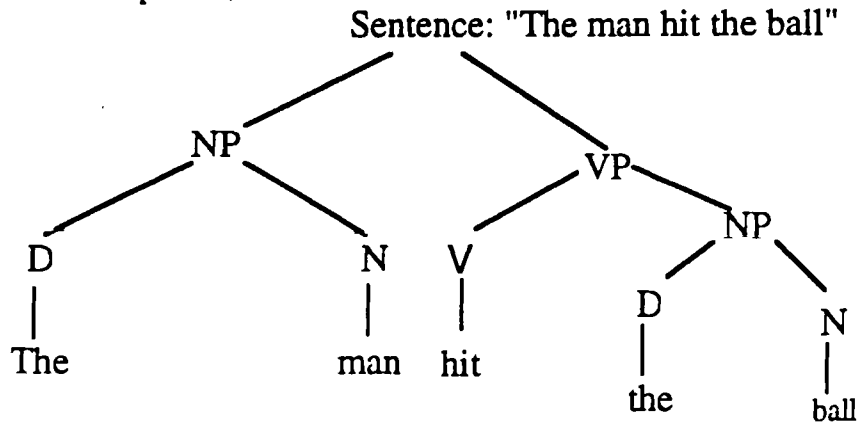

Figure 7

Type 1 or context-sensitive grammar is a phrase structure grammar that satisfies the condition that for any proposition "*p* -> *q*", *q* has the same number, or a greater number, of symbols as *p*. The language generated by this grammar is known as a context-sensitive language. These languages are not very suitable for the formalisation of statements that have grammatical constraints. For these it is perhaps better to return to a context-free grammar that is already equipped with added constraints.

A context sensitive language is, however, very useful for the representation of propositions in a more complex natural language. The sorts of propositions I am thinking of are those where the context in which the phrase is used becomes important, or where the sentence is already heavily embedded. Indeed any instance of ambiguity might render the sentence inexpressible without the use of a more comprehensive grammar.

For example, in the proposition *"xwz -> xyz"*, *y* only maps on to *w* where the conditions of *x* and *z* are also met in exactly the same way as they are in the first half of the proposition, *xwz*. So that *A* is in a particular context between *x* and *z* is seen to be of great significance. Thus, if *y* is also between *x* and *z*, with *x* still placed to the left of *y* and *z* still to its right, and *x* and *z* have the same meaning that they did in *xwz*, then *A* and *y* can be said to be equal. So for: *x* = 'Mr.Bun', *w* = '(the baker)' and *z* = 'bakes cakes', *y* also represents '(the baker)' where *x* and *z* consistently stand for 'Mr.Bun' and 'bakes cakes', respectively. Thus, it can be seen that *w*'s context is very important and only something that a context-sensitive grammar can recognise and interpret.

The grammar that is the most complex and therefore in the highest position on Chomsky's hierarchy is Type *0*. It is a grammar that can generate an unrestricted set of grammars and with every set being recursive Type *0* is well able to parse sentences of English in context. However, it is a grammar that has a mainly theoretical application for it is most commonly used for examining the complexity of a particular computation to see if that computation can be generated by any of the other grammars.

In any of the other grammars, types *1* to *3*, the number of non-terminal symbols on the left-hand side of the implication sign has to be equal or less than the number of terminal symbols on the right-hand side. In the type *0* grammar there can be any number whatsoever of non-terminal and terminal symbols. So with no correlation being necessary there are no constraints of any kind placed on this type of grammar.

I shall turn now to the machines that are capable of recognising different phrase structure grammars. This section will explain what they are and the basics of how they operate so that we can see just how they set about recognising a grammar. I will discuss the corresponding grammars as I go along.

5.2.2. The machines and their behavioural properties

Finite State Machine (FSM) and Type 3 grammar

The simplest form of machine is called a *basic machine*. It is described as a *combinatorial* machine because it is only able to interpret a set of input information and from that produce a set of output data that is a combination or function of the inputs. The difference between a finite state machine and a basic machine is that the finite state machine is a basic machine with the improved capability of an internal state that alters in relation to the input. So that the output of a FSM (which is already specified) is a function of the input and its internal state. Both machines have very limited capabilities.

Provided we know the initial state, the input and the transition function, the behaviour of the FSM can be determined absolutely. With the same information it is also possible to specify the set of all final states of this machine. As a result of this increased power the FSM is now able to recognise grammatical sequences as being members of a specified grammatical set. By 'recognition' I mean that the machine will react in one predicted way if the sequence is a member of the set or in another different way, again predictable, if it is not.

An FSM is considered to be deterministic if given a specific state *s* the same input symbol will always cause the FSM to move into a particular state and no other. However, there is a non-deterministic FSM (NFSM), which can move into more than one possible state on receipt of the same input symbol. It has more than one possible transition and because no weights can be assigned to these transitions it can be described as a *possibilistic* machine.

The FSM is capable of recognising only the type 3 grammar that is identical with the regular expression language, and is known as the regular or finite state language. Indeed both the FSM and the NFSM can accept the same sets of words in the type 3 grammar. One significant advantage of the NFSM is that it can be a smaller machine

since the transition state can remain unspecified. Extra space is required in the FSM because the 'go to' state has to be stipulated since it is a deterministic machine. When the state outcome is not important a NFSM is used.

Push Down Machines (PDM), Non-deterministic Push Down Machines (NPDM) and Type 2 grammar

The structure and capabilities of the PDM (here assumed to be determined) dictate that it lies somewhere between the Turing Machine (TM - see below) and a FSM. A TM is a FSM with an infinite auxiliary memory in which information can be stored and recalled in any way at all. The difference between the TM and the PDM is that in a PDM there is a restriction on the storing and recalling of information in the auxiliary memory. In this way the restriction resembles a stack where the object or symbol that is last-in can be picked off first or the one that is first-in can be picked off last. The implication of this is that symbols are always stored at, or recalled from, the top of the stack. When a new symbol is added to the stack it pushes the previous symbol that was put there first down one place in the stack, so that the first symbol is now in the second place in the stack.

The PDM is made up of an input tape, a FSM and a stack. The stack is its memory which can be compared to random access memory, (RAM). As we have seen the FSM does not have any memory so that the addition of the stack or memory to the PDM increases the capabilities of the machine. An added capability that a PDM has over a FSM is that it can recognise the class or irregular sets of context-free grammar. It is this class that contains regular or finite state languages and is therefore of great value for the generation and translation of computer languages.

The stack is represented as a string of symbols from an alphabet, and because the stack is assumed to be arbitrarily long any number of symbols can be added to the top of the stack. The furthermost symbol to the left is considered to be the first in the stack. When a symbol is added to the stack it is called 'push' or 'load', and when a symbol is

deleted from the top of the stack it is known as 'popping'. The stack is non-uniform because only the top is added to or taken-away from. This type of stack memory has two advantages, firstly it has no addressing scheme, and secondly, only two commands are needed for the storing and recalling of information. These commands are "push" and "pop".

We have seen that in a deterministic push down machine the output is determined by whatever the specific input is. This is not the case for a non-deterministic push down machine. For instance, the PDM has one possible internal state to which it can move on receipt of an input, whereas, for the same input, the NPDM has a number of different possible states it can go to. So the next state is not determined or determinable. It might well happen that for the same input both machines output the same state but, because only one response can be predicted it cannot be stated categorically that this will happen on every occasion.

Being non-deterministic does not necessarily mean that a new class of states is added to the system, for the same states may be present in both machines. Being non-deterministic may even mean that the overall system of the NPDM is smaller than the PDM because the latter has a special set of output states that are necessarily present.

Linearly Bounded Turing Machine and Type 1 grammar

The Linearly Bounded Turing Machine (LBTM) is a machine that is similar to the TM except that it has a limited amount of tape that contains only the input string plus an extra two squares that are to hold the end markers. This limitation means that the machine is restricted in its power to recognize some symbol strings. However, even when the length of tape is increased as a linear function of the length of the input string the computational ability of the machine remains unaltered because no additional information is being added in the form of new symbol strings. This machine, a linearly bounded memory machine, is capable of recognising the Type 1, context-sensitive

languages. The tape bounding means that the machine is only capable of generating and interpreting a particular set of input strings or symbols in a limited direction on either side of the tape head. This 'limited set of symbols' between the two 'bounds' is the context to which the 'read/write head' is sensitive, in this case it is from 'alpha' to 'kappa'.

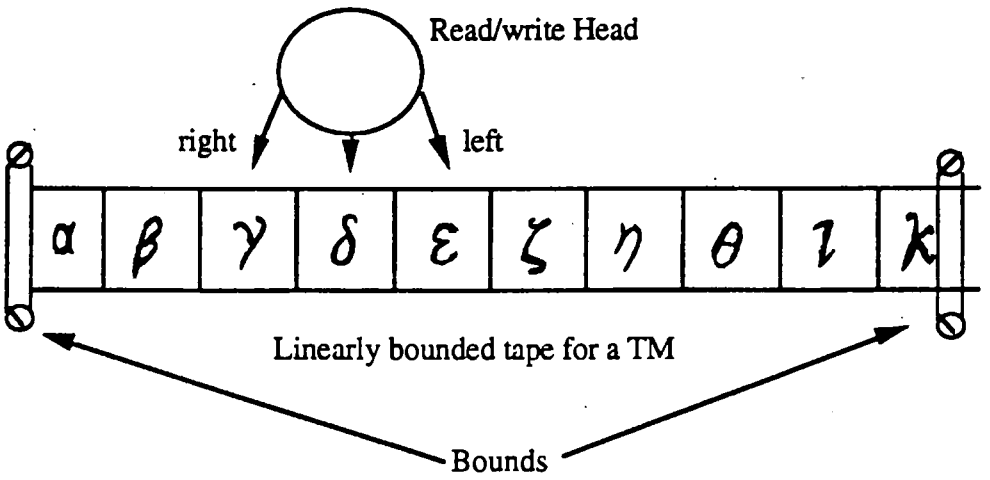

Figure 8

Turing Machines (TM) and Type 0 grammar

A Turing Machine is a Finite State Machine that has an infinite auxiliary memory in which information can be stored or recalled in any manner by the movement of the tape in either direction (right or left), and by an unspecified number of squares. The basic hardware of a TM is in two parts; a head and a potentially infinite tape. The head can read or write a symbol, move left or right or stay put in relation to the cells or squares marked-off on the tape. The tape is of an infinite length and it extends from each side of the head. It is marked into square cells that can contain symbols from an alphabet set written in. A machine of this sort must be capable of accomplishing a number of simple tasks. Those tasks are firstly, that it must be capable of changing the symbol on one of the observed squares, and secondly, it must also be capable of changing one of the observed squares to another square.

The symbols on the tape are formed from a finite set of symbols called the external alphabet or "*I*", that consists of a lowercase English alphabet, Arabic numerals, punctuation marks and a symbol for a blank. By reading from and writing to the cells of the tape the machine communicates with the outside world. This communicative ability is demonstrated by the movement of the head to the left, or to the right or by its remaining static.

It is only possible for the machine to reside in any one of a finite set of states, *S*. These states are indicated by the use of the lowercase Greek alphabet in the diagram of the LBTM. The transfer of the machine from one instruction to another can be seen as equivalent to a change in the 'state of mind' or internal state of the machine.

The TM has three main functions, Machine Function (MAF), State Function (STF) and Direction Function (DIF). The resulting TM computations are simply a matter of executing and repeating the actions of the MAF, STF, and DIF. At any given time, the machine state plus the content of the scanned square will either cause the machine to take action (moving right or left) or to halt. If it reacts at all it will be to perform three actions before the next appropriate time interval. The actions are, (i) that the square being read is erased and another symbol is printed on the square (MAF), (ii) that the internal state is changed (STF), and (iii) that the head moves to the left or to the right or remains static (DIF).

Every Type 0 language generates a recursively enumerable set of languages that are made up of arbitrary sets of symbol strings. In terms of the grammar already set out above a Turing Machine could be constructed that could recognise and successfully parse its sentences. No restrictions or constraints exist on the production rules of this language.

I will now give a resumé of what has been said in this section. Then in the next section I will take a look at Dretske's stratification of intentionality into three levels and

the implications that this has on the related capabilities of the systems he proposes as capable of the different levels.

5.2.3. A resumé of Chomsky's hierarchical stratification

As a formal symbol system a machine is capable of recognizing and generating a specific set or class of languages. The level of its capability will depend upon its internal states or architecture and its auxiliary memory that stores and retrieves input information. Being a very basic machine the FSM has no auxiliary memory so it is restricted to the generation and acceptance of regular grammars and languages. Languages of this sort are very primitive but they can be used to implement things such as text editing and command languages. The PDM and NPDM are slightly more complex with an auxiliary memory but because there is a restriction on their capacity to store and recall information, they can only accept the class of context-free languages. The linearly bounded TM is similar to TM in every way except that it has a restricted tape which means there is a limit placed on the input strings that it finds it possible to recognize. A result of this is that it can only accept, recognise and interpret context-sensitive languages. The TM, with its infinite auxiliary memory, can generate unrestricted grammars and information can be stored and recalled in any manner at all.

It is possible to see from this fairly straight-forward hierarchical arrangement that a relationship exists between the architecture, or internal states, of a machine plus the input it is capable of receiving from its environment and the capabilities of the machine to carry out certain tasks. The tasks are 'certain' because in Chomsky's example they are set out for us and consist of the recognition and interpretation of phrase structure grammars that themselves vary in complexity from the most simple, type 3, to the most complex, type 0.

The architecture and the environment can be seen to have had a substantial influence on the capabilities of the systems in question so that we can see that the simple

machine, with only the most limited connection to the world, is capable of only the simplest tasks. The most complex machine with a number of obvious links to the world is capable of generating and recognising the most complex grammars. And in between the two there are two machines, the one that is slightly less complex in design than the TM, the bounded TM, with constraints on its tape which mean it has a more limited access to the world. It is less capable than the TM but more capable that the FSM, the PDM or the NPDM. And finally, the deterministic and non-deterministic push-down machines that have a more complex design than the FSM because of their memory facility but with the constraint of having no tape which means they are unable to recognise context sensitive grammars; because of this they are more capable than the FSM but less capable than the linearly bounded TM and the TM.

Chomsky deals with machines and machines states and their relation to the world is shown through the languages they can use and the phrase structure grammars they can generate, recognise and interpret. He does not deal with mental states, nevertheless he shows that there is a tenable relation between the structure of a system, its link to its environment or domain and the things of which it is capable.

Dretske shows this relation by looking at the possibility of a number of different types of system, organic and inorganic, having particular mental states. This possibility is based on his notion of dividing intentionality into three levels and then examining which level of intentionality different systems exhibit depending on their capabilities. We shall see that the idea is that a simple system, capable of only simple information processing and first level intentionality, can occupy only the most basic of mental states. A more complicated system, that can exhibit some understanding of its incoming information is capable of a higher level intentionality and can therefore possess higher level mental states.

Having already looked at Dretske's theory at some length in a previous chapter, section 2.8, I will briefly go over the main points that were made there and then go on to show the

logical formulation with which he forms his divisions of intentionality. Following this I will explain how his division demonstrates that a relationship can be seen to exist between the system, the stimuli it can respond to in its environment, and its capability to process the information it receives. However, I will also show that if this relationship is shown in a hierarchical stratification it is bound to fail because it deals with mental states that cannot be differentiated in the way that machine states can be.

5.3. Dretske's hierarchy of intentionality[5]

Dretske outlines three levels of intentionality and he attempts to relate these three levels of intentionality to the plasticity or flexibility of a variety of systems to extract information from their incoming perceptual signals. He concludes that only systems that are capable of reaching third order intentionality are flexible enough to be able to completely digitalise information and disclose the semantic content contained therein. The human mind is the most effective system for reaching this level, but it is not the only one that Dretske believes to be capable of this level of intentional behaviour.

The human system has the plasticity to digitalise the nested analogue information that it perceives in a signal and from that information extract the semantic content. The semantic content that it extracts will be the one which is most relevant to it and the belief system that it already has. The implication being that if two people have the same sensory input they might still each extract a different semantic content[6] and it all depends upon what is of most interest to them. However, the main point is that human beings can form beliefs, and it is this that distinguishes them as cognitive systems from thermostats which are mere information processors. So, it would seem that for any system to be capable of forming beliefs it would first need to be capable of processing its incoming information at a level of third order intentionality.

5.3.1. Intentional states and levels of intentionality

As has already been stated all information-processing systems can occupy intentional states of one order or another, so that some systems are capable only of information processing, whilst others can understand and conceptualise the information they receive. A physical state carries information about a source, which is to say that it occupies an intentional state relative to that source. If we take 'S' to stand for the signal that the system receives from the source of the information, then in any of the orders of intentionality S states that 't is F', but it does not necessarily state that 't is G', regardless of the fact that anything that is F is G. The information in structure S has a propositional content that possesses intentional characteristics.

Dretske gives his orders of intentionality as follows:

(1) First Order of Intentionality - (Contingent)

(a) All F's are G

(b) S has the content that t is F is G

(c) S does not have the content that t is G

The signal 'S' has a content that exhibits first order intentionality. All information-processing systems exhibit this order of intentionality for it depends solely on the interaction of the system with its immediate environment. The best explanation of first order intentionality is that it is possible to receive some information about a thing without receiving all of the information about it. So, for example, a thermostat receives the information that the room temperature is too high but it has receives no information about why this state of affairs has come about. The thermostat can only receive information of a particular kind from a general information signal for nothing else is of relevance to its successful operation.

(2) Second Order of Intentionality - (Natural)

(a) It is a natural law that F's are G's

(b) S has the content that t is F

(c) S does not have the content that t is G

At this level it is not the signal,'S', but the system that exhibits the information content of second order intentionality. In this instance it is possible to know something without necessarily being aware of all its underlying implications. For instance, it is possible to know that a pool of water is freezing without knowing that the water is also expanding, even though it is a natural law that water cannot freeze without expanding. The notion of its expansion is a piece of implicit information that depends upon the natural laws that hold in the empirical world.

(3) Third Order of Intentionality - (Necessary)

(a) It is analytically necessary that F's be G

(b) S has the content that t is F

(c) S does not have the content that t is G

Again the system, and not the signal, exhibits third level intentionality. For example, it is possible to believe that 12 is the number you get when you multiply 3 by 4 without necessarily knowing that 12 is also the sum of 7 and 5. Knowing one does not entail knowing the other, nor does it rule it out. Thus it is possible to know something is the case without knowing all that there is to know about it. Dretske's own example is that we can know that the solution to a mathematical problem is 23 without being aware that 23 is also the cube root of 12,167; where t is F is '23' and t is G is 'the cube root of 12,167'. So that just because it is necessarily so does not make it also necessary for us to know it.

It is this third order that I am most interested in for having the capability to reach third level intentionality means that a system can be seen to possess the flexibility to do all manner of things, from ignoring some pieces of information whilst selecting others, to the extraction of the relevant semantic content and the formation of appropriate beliefs.

5.3.2. Semantic, or propositional, content

If we cast our minds back to chapter two I showed there a diagram of three concentric rings[7] that Dretske uses to show what he means by analogue and digital information, and how we digitalise information to form beliefs. According to Dretske belief corresponds to the outermost informational shell, for a formed belief is the only completely digitalised piece of information and all other information in the signal remains nested in analogue form inside the outermost shell. A visual experience is in analogue form and only the information that is selected and conceptualised becomes digital. Once a system has reached the level of belief it must be sure that it has a semantic content that is commensurate with the formation of what would be, in its own personal context, a true belief.

Dretske describes this semantical content as the propositional content that demonstrates third order intentionality. A belief, unlike an information structure, has an exclusive propositional content which the system has formed. This accounts for the belief that a system forms not being in any way determinate. An information signal, on the other hand, carries all the nested information that is possible within that one signal. That a human being can have a belief about *X*, and therefore also an understanding of its semantic content, makes that belief distinct from the beliefs about *X* that just exist *per se*. The beliefs that exist *per se* are those that are implicit in the incoming information but which are not of relevance to the perceiver at that time.

For example, the statement 'I am sitting' conveys more information than just that I am in a sedentary position. It also gives us lots of negative information, for example, that I am not running, standing, or swimming. However, for the listener the semantic content of my utterance is simply one piece of the whole thing, and perhaps for them it means only that I am sitting. That the listener can extract the meaning from the information that I convey, means also that they are able to form beliefs about that

information, thus meaning and belief formation have, for Dretske, the same level of intentionality. For Dretske if we can understand 'meaning' we can also understand the processes behind forming and holding beliefs.

From this account and the one in chapter two it is possible to see that in Dretske's hierarchy knowledge and belief have a higher order of intentionality than just being able to process incoming information. It is also clear that what we believe and the beliefs themselves are quite distinct even though their contents are often logically interdependent for it is only the beliefs that we have formed that show that we have an understanding of the information that we have received through signals from our environment. Dretske would say that only through the formation of true beliefs, or beliefs that are appropriate to our circumstances, can it be seen that we have successfully selected the most relevant piece of information from the incoming signal, stripped away the unnecessary pieces, and extracted the semantic content. Then, and only then, can I show that I have been capable of completely digitalising some selected piece of my incoming information.

5.3.3. Systems, environments and capabilities according to Dretske

Being able to reach only a first order intentionality is equated with any system that is capable of no more than processing information. These are simple systems with a very limited environment and only the capability to process information; they have no knowledge or understanding of the information with which they are dealing. Second order intentionality is associated with a system's capability to know something of the information it processes. The system, although still relatively simple, is capable of possessing epistemic states, but not of fully understanding that information. Finally, a third order intentionality is only achievable by systems that can process their incoming information, ignore some pieces of it and select others from which they can then extract the semantic concept that is most relevant to them and form beliefs. These 'belief-

forming' systems are also capable of adapting their behaviour to suit their new beliefs, and altering their perception of events, past, present or future, to accommodate their new or changed concepts.

It is possible to make a hierarchical stratification of Dretske's division in much the same form as Chomsky's, and this is done in the diagram below. In a similar way to Chomsky's the system that is capable of the top level of behaviour is also capable, as a matter of course, of accomplishing the tasks at the lower levels. So that anything that is capable of forming beliefs about selected pieces of information that it perceives is also capable of processing information, knowing what information it is and understanding it in full. A Turing Machine with no linearly bounded tape is capable of generating unrestricted grammars, but also of generating context-sensitive grammars, context-free grammars and primitive grammars. In Dretske's hierarchy there are systems that are only capable of processing information and nothing more, similarly in Chomsky's hierarchy the Finite State Machines can generate only primitive grammars and nothing more.

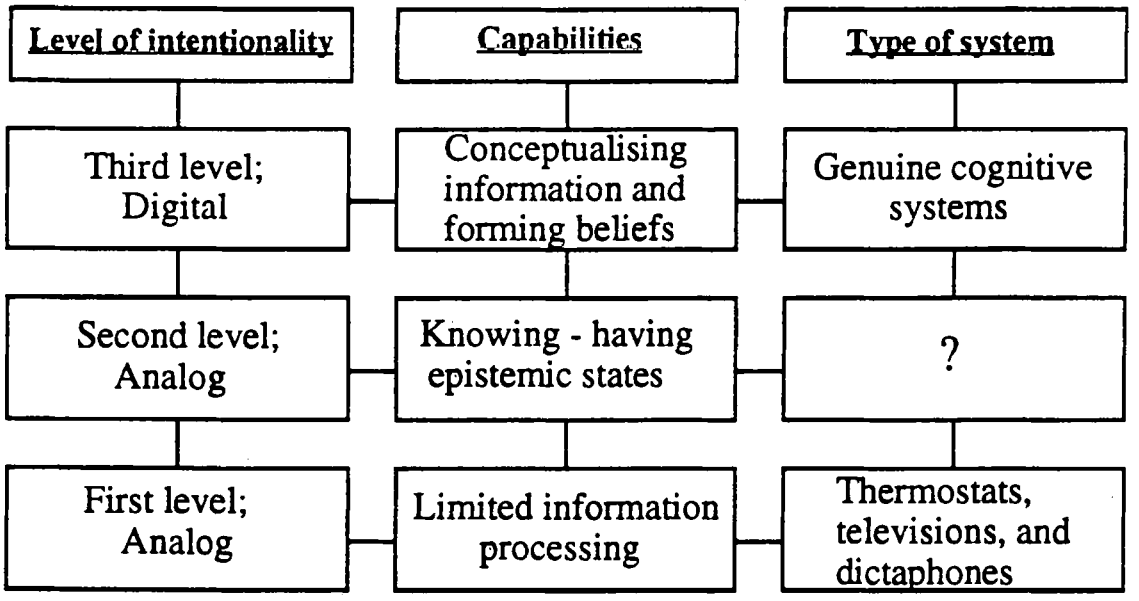

Figure 9

Both hierarchies are successful at showing that a relationship exists between a system's capabilities and its complexity of design and the extent of its domain,[8] but in the next section I will put forward an argument to show that Dretske's stratification of mental states is prone to failure whereas Chomsky's hierarchy because it is about machine states, is not at risk in the same way. The following section contains four criticisms that I shall make concerning Dretske's proposal for a hierarchy of intentionality. The last two of these criticisms deal directly with why Dretske's hierarchy is unsuccessful.

5.4. A criticism of Dretske's work

The first problem lies with Dretske's concentric ring diagrams which I believe do not convey the information that he expects. Secondly, Dretske's use of the terms 'analogue' and 'digital' is suspicious because he offers at least two incompatible senses of 'digital' and uses them synonymously. The third problem is that for levels one and three Dretske offers suggestions for the type of system that would best fulfil the capabilities, but for the second, or nomic, level there is no such possible system. Finally, Dretske claims that 'frogs, humans, and perhaps some computers' are 'genuine cognitive systems' capable of third level intentionality; and this is simply misleading.

5.4.1. Faulty diagrams

In this first criticism I shall argue that his diagrams are both logically and intuitively problematic. They confuse logically because they go against the Venn diagram conventions which are logical representations of sets of states of affairs. In a Venn diagram of logical implication 'P then Q' becomes a large circle 'Q' with a smaller circle 'P' contained in it, so that everything that is P is also Q. At the same time this states implicitly that not everything that is Q is also P.

The logical implication of if P then Q / P -> Q

Anything that is P is also Q. The converse does not hold; for it is not the case that anything in Q is also in P unless Q -> P is also implied. This would mean that the relation of P and Q is one of equivalence and not one of just implication.

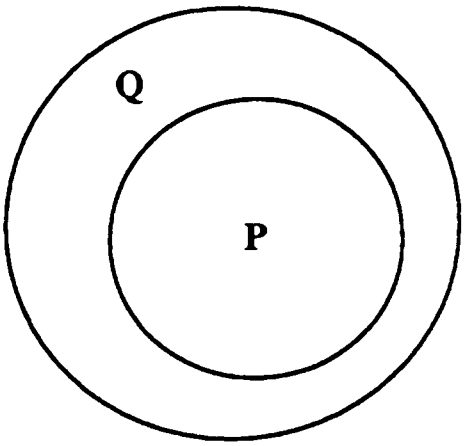

Figure 10

In the following diagram[9] Dretske's use of Venn-like diagram suggests that everything that is a quadrilateral is also a square; but this cannot be so for a quadrilateral might be a trapezoid or a rhombus, but it does not necessarily have to be a square. The problem is that to reach embedded information one would intuitively follow a natural progression inwards from the general informational signal to a particular piece of information, but Dretske's diagram appears to work the other way from the signal, marked as 'S', as a whole outwards through the analogue information and finally to a piece of information which is completely digitalised.

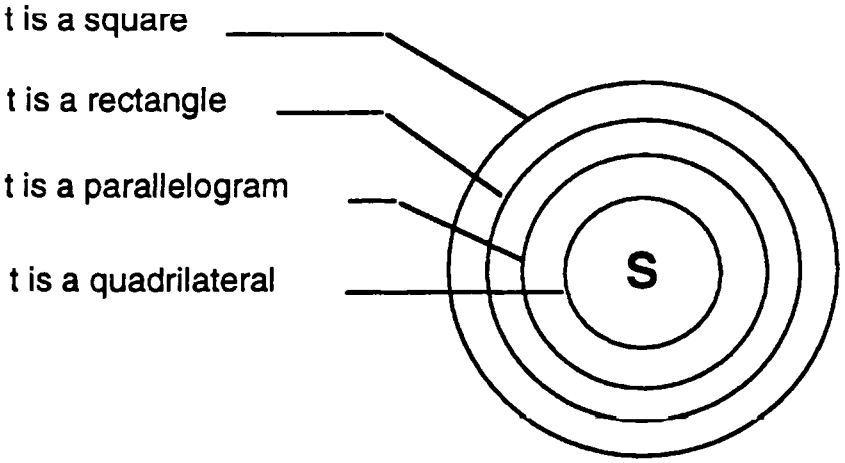

Figure 11

It seems then that his diagram is 'back to front' because if we talk of one piece of information as being embedded in another piece of information then the obvious thing

would be for the most specific piece of information to be at the middle of the diagram which would be followed by a progressive movement outwards towards less and less specific information that is analogue in nature. This would mean that the diagram is reversed and the analogue information of structure 'S' should be the outermost ring.

However, a problem arises here because Dretske explains that in the process of digitalisation the extraneous information is stripped away from the most particular piece that has been selected and we are left with the required concept, but, if the piece of information that we have selected and examined is the most particular piece there cannot be any more spurious information to strip from it.

That human beings are capable of making allowances for the message carrier is something that distinguishes them from voltmeters. Human beings have the selective capability to disregard the carrier of the information and also to perceive what level of influence the carrier has had over the message, and then they are capable of extracting the carrier and its influence from the message and finally leaving what is the most relevant or specific piece of information for them. The human understanding of information is represented by the outermost ring in this diagram yet Dretske states that as we understand we extract the semantic content and information is lost; why then is the outermost ring the largest and the ring within which all other information is stored, thus suggesting that no information is in fact lost.

Another problem is Dretske's use of the term 'embedded' when he speaks of two or more pieces of analytic information. The difficulty is simply that two pieces of information that are logically equivalent cannot be embedded one inside the other. Again this can be demonstrated using venn diagrams, (see Figure 12). The best approach to this problem is to begin by stating three of the definitions that best define an analytic truth[10]:

1. the concept of the predicate is contained in the concept of the subject, (Kant).

2. it is possible to prove or disprove an analytic truth or falsehood by means of the definitions of logical laws, (Logical Positivists).

3. a statement is an analytic truth if it is true in virtue of the meanings of its constituent terms.

If the third definition is used with regard to the third level of intentionality, then an analytic relation is one of equivalence where both terms are intersubstitutive, *salve veritate*. In this case then the definition of one term can be swapped with the definition of the other, equivalent term. In Frege's example, 'The Evening Star is the Morning Star' it would make no difference to the sense, the reference or the truth value if I were to say, 'The Morning Star is the Evening Star'. Only the order of this statement has been altered by this reverse construction. Dretske's 'analytic' example would appear to be wrong in this instance because to be intersubstitutional both pieces of information have to be equivalent and with their being equivalent if follows that neither piece can be embedded in the other. Neither piece of information can be more specific or more unique than any other piece since there can be no use of comparatives in relations of equivalence.

I think Dretske would argue here that what is more significant is the notion of 'relevancy', that is, the person selects a particular piece of information and conceptualises it, thus forming beliefs about it and this most relevant piece of information can be equivalent to another piece but the difference is that it is not relevant to that person. But this still leaves Dretske with the problem of diagrams that do not accurately represent what he wishes to say. In the diagram on the left the outer most ring is meant to represent the most specific piece of information that has been embedded in the initial structure 'S'; it stands in an analytic relation to other information.

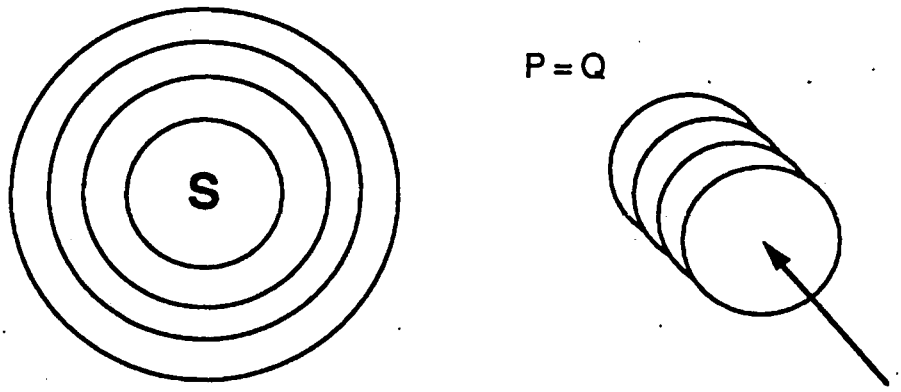

Figure 12

This suggests that the diagram would have to look somewhat different. I propose that each of the circles would have to overlap exactly if the information represented is logically equivalent in any of the accepted senses of analytic, or semantically equivalent in Dretske's sense. There could be no embedding of equivalent pieces of information one inside the other. The diagram, I would suggest, would have circles of equal diameter which would represent their equivalence of information content. The unfortunate consequence of this is that all the pieces of information would then look like one circle, as they would if the second diagram were looked at along the direction of the arrow.

To summarise, my objection to Dretske is that two equivalent things cannot be nested one in the other, and I believe that Dretske's use is misleading in both an intuitive sense and a logical sense. It does not seem possible for the reader to infer from Dretske's diagram of analytically nested circles only that the information in each circle is definitionally equivalent. This would need to be stated explicitly. In the same sense I believe it is misleading for him to confuse the notion of 'embedding' with the notion of being 'inside', for equivalent pieces of information can be 'embedded' by being logically definitionally equivalent without one piece necessarily being inside the other. This notion is especially confusing when used in the context of analytic relations, as Dretske has done.

Now I shall move on to offer a criticism of Dretske's use of 'analogue' and 'digital' which is, to say the least, idiosyncratic. The common use of the analogue/digital distinction is to signify a difference in the way that information is carried about particular properties. These properties can vary, so for the example of 'pressure' the information about it is carried using a barometer, and for the example of 'temperature' the information conduit is a thermostat. Dretske bases his use on this distinction, but, as he himself says, in a 'slightly unorthodox way'.

5.4.2. Digital and analogue

Dretske is not concerned with how the information being carried is encoded, but rather how the facts or information about variable properties, such as pressure and temperature, is represented. His information-theoretic use of the distinction can be said to mark 'the different way facts can be represented'.

A signal is said to carry the information that 's is F' in digital form if there is no other information carried in the signal. More precisely, what is meant is that there is no other information that is also embedded in the s's being F. Any other information that is carried in the signal, but not that which is already embedded in s's being F, is said to be carried in analogue form and all signals carry information in both analogue and digital form. It is true of every signal that it carries more information in analogue than in digital form, and in the move from analogue to digital information a lot of peripheral information is necessarily lost. The information that is carried in digital form is then, the 'most specific, most determinate, piece of information' that the signal carries. It is the semantic content of the signal and the only piece of information that is carried in digital form. Everything else is carried in analogue form.

Dretske illustrates his version of this distinction with the communication of a piece of information about a cup of coffee. The statement 'The cup has coffee in it' tells us only the most specific piece of information that there is coffee in the cup. This statement

expresses '*all* the information a signal carries' and is represented in digital form. An analogue representation of the fact that there is coffee in the cup might be a photograph of the scene. In the format of the photograph there is much more information available and no one piece is any more specific than any other piece.

Thus, for Dretske, a statement is a digital representation of the information being carried in a signal and a picture is an analogue representation of the same signal. 'The information a picture carries in digital form can be rendered only by some enormously complex sentence, a sentence that describes every detail of the situation about which the picture carries information.' The old adage 'a picture is worth a thousand words' is very significant for Dretske, for it conveys his argument very clearly. It would need to be a very complex sentence indeed if it were to adequately describe the state of affairs in the picture. Dretske argues that what usually happens when we describe a scene is that we convey all the analogue information because the digital information is much more specifically what the scene would mean to me. The semantic content is that particular piece of information most relevant to the person looking at the scene or the piece that they extract from the verbal description that I give them. But it might be argued, and I believe more reasonably, that whenever I describe a scene to someone I will give them only that information that has seemed relevant to me, that is information that has been digitalised. For if Tom were to describe the same scene to me he would give me different information, that is the information that he has in digital form that seemed most relevant to him. Similarly witnesses to an accident will always give differing accounts of the events that led up to the accident for they see things from their own, unique perspective.

Even if we accept that Dretske's use of the terms 'digital' and 'analogue' is 'slightly unorthodox' his use in this context remains misleading because he seems to want to mean two things simultaneously. In one reading 'digital' means 'all' and in another it means 'most particular'. 'Digital' in its most common usage means 'discrete packets' of

information and Dretske's use could conceivably be interpreted in this sense when he says that it is the most specific piece of information that is digitalised. By selection and categorisation a particular piece of incoming information is extracted for attention, it is this piece that is the 'most particular' or specific and it is this piece that counts as the overall semantic content of the signal as a whole.

However, elsewhere Dretske states that the outermost ring is 'all the information carried by the incoming signal'[11] and that to form a concept we have to move inwards and strip away any irrelevant information; so it would then seem as though the outermost ring cannot be the most specific piece of information after all since it must, at least, be representative of the most general information from which the semantic content is extracted.

If more than one piece of information is carried in digital form then more than one piece could be the semantic content of the informational structure. This in turn suggests that the semantic content is not, in fact, unique as Dretske has argued. This means that the semantic structure is not the information that is carried in digital form, but that piece of information that has been *completely digitalised* and this then represents the *outermost informational shell* "in which *all* other information is nested (either nomically or analytically)".

The distinction that Dretske sets up might be better thought of as being between 'digital' and 'completely digitalised', these are the 'all' and the 'most unique' pieces of information respectively. But this does not seem to be entirely plausible either since only a part of the information within the outermost informational shell is carried in digital form. The other part or parts are carried in analogue form and to get to the most specific piece of information we have to move outwards through the analogue information and the information that is stored in nomic and analytic form; but then to conceptualise that piece of selected information we have to move back inwards again

through all the information that is nested there and choose what is relevant in our own specific circumstance.

But this is not all, for there is still greater confusion surrounding his use of 'digital'. I shall explain. The idea of digital information is carried through the whole chapter as being the outermost informational shell and the semantic structure of the signal; but, just three pages from the end of the chapter there is a dramatic change which states that the semantic content is now the part of the signal that has been completely digitalised. So the outermost ring is no longer equivalent to digital information. The change was required because the definition of 'semantic structure' needed 'tightening up' since it was possible for more than one piece of information to be carried in digital form, in which case they would be analytic and nested within the outermost informational shell.

My third criticism of Dretske's attempt to stratify intentionality is this; Dretske establishes three levels of intentionality and for the first and the third he offers a variety of systems that are capable of achieving each of the two levels, but he offers no systems that are capable of second level intentionality. The systems that are capable of third level intentionality are also capable of first and second level intentionality, and those that are capable of first level intentionality are capable of only that, and nothing more. So presumably those systems that are capable of second level intentionality could have epistemic states and also be able to process information, yet still not be able to form beliefs about that information. I shall now examine the implications of the missing second level systems, the missing link in his chain of three levels of intentionality.

5.4.3. No systems equate with second level intentionality

The three distinct forms of intentionality are based on the amount and the extent to which information can be processed by different systems, but Dretske does not offer systems that correspond to each of the three levels. At the first level he proposes that

simple mechanisms, such as a television, a dictaphone, or any conduit of information, can process information. At the third level he claims that frogs, human beings and perhaps some complex computers, are among those things that can cope with the elaborate process of selecting some pieces of information over others, extracting the semantic content from this perceived information and finally forming beliefs about it. However, he seems unable to posit any systems that can achieve the second level of intentionality but not the third. This poses a fundamental problem with the divisions he has drawn up. Either it would be a good idea to have only two levels of intentionality and equate level two with level three, or it would be advisable to have systems that are capable of more than level one but less than level three, and thereby justify the existence of systems that can have epistemic states but not go as far as to form beliefs about them.[12]

There are two possible reasons for Dretske's hesitancy in citing a system capable of second, but not third level intentionality. The first reason is that it is easier to say what a system can do rather than what it cannot, and the second reason is that it is very difficult, if not indeed impossible, to differentiate between different mental states because they are not things that are finite and measurable. I shall now look at each of these reasons in more detail.

Negative claims are difficult to make

In chapters three and four I examined the basis on which it is possible to decide what capabilities a system possesses and how we then set about ascribing mental states to that system. Drawing a limit to the capabilities of a system is a simple matter when dealing with systems such as thermostats for their capabilities are simple and refreshingly obvious. They can detect that the surrounding temperature is too hot, too cold or that it is just right, in which case it receives no signal from the environment. They respond by switching the heating system off, on or by remaining static. They are

simple information processing units of which it is possible to say precisely those things of which it is capable and those things of which it is incapable. For instance, it is not capable of making tea, chatting about the weather or grooming the dog.

But the matter of what a system is and is not capable becomes increasingly more complex as the system itself becomes more complex both internally and in relation to its environment, that is, its perceivable domain extends so that those things it can respond to increase in number and variety. Thus when we come to examine the behaviour of, for example, a frog or a cat any judgement about its range of capabilities is going to be quite problematic. Observation and experience of my cat's behaviour tells me that when it miaows and rubs my legs with its head it wants food; this is not difficult, nor very interesting behaviour to attempt to interpret. However, my cat seems to have learnt other activities, such as how to get me out of bed in the morning or lead me to the cupboard where the food is kept. What makes these more interesting behaviours is the question of whether or not the cat actually knows that pushing things off the bedside table will get me up, or that if it leads me to the cupboard I will know it wants food. It may be the case that my cat has made some sort of connection with me being vertical and ambulatory and its being fed. If we accept that the cat's behaviour is an exhibition of knowing behaviour, that is, that the cat does have epistemic states, how then is it possible to draw a distinction between its knowing behaviour and what would count as believing behaviour.

We have seen how difficult it is to ever tell when another human being possesses a state of one sort or another, even though in the case of other human beings we have a shared language with which we can speak of our mental states and offer confirmation or denial of any state that is attributed to us. How much more difficult it is then to tell that a system other than a human being, that does not have the shared human language, is occupying a particular mental state. All we have to go on in the case of any non-human system is its behaviour and what we know of its physiology and

neurophysiology. For some of the systems in Dretske's division of intentionality it would seem easier to give them the benefit of the doubt and make the positive claim that they are capable of third level intentionality, in which is subsumed the second level, than to say negatively of a system that it is capable of knowing things about its world but not of believing those same things.

Indeed 'knowledge' has a definition as 'justified true belief'[13], and in this sense knowing is also believing even if the individual does not actually go through the process of thinking, 'I know that it is raining, therefore, I also believe that it is raining'.[14] This definition cannot be inverted for I do not know everything that I believe. A simple example is just to turn the earlier proposition around so that it reads as, 'I believe it is raining, therefore, I also know that it is raining'. It is easy to see that this is inconsistent for knowledge claims are definitive, based on conclusive evidence and emphatic, beliefs, on the other hand, are often based on the flimsiest evidence because we want something to be the case and we will accept the first thing we find to back our belief up; often described as 'clutching at straws'. Saying 'I believe' leaves one open to accepting the converse if it is proved, whereas saying 'I know' suggests that your mind is made up and no new evidence will shift your opinion.

This can be shown using Venn-like diagrams:

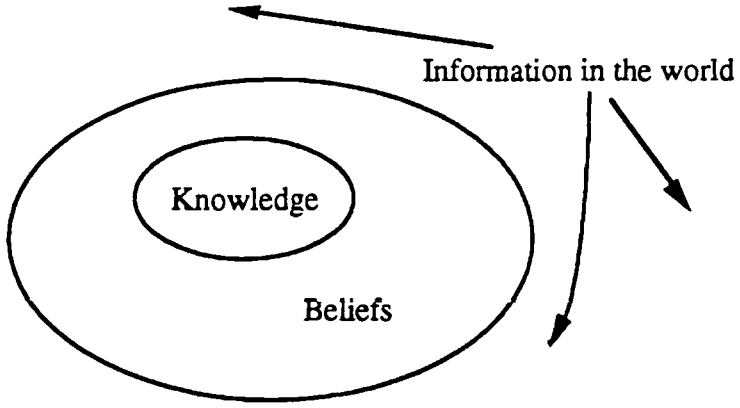

Figure 13

However, it must be asserted that the rings that represent these mental states do not have fixed borders except that the circle of 'knowing' must always be inside a ring of 'beliefs' or 'believing'. Outside the ring of 'beliefs' there will be more beliefs that are irrelevant to this particular piece of knowledge or in the form of information that have yet to be encountered.

That it is easier to make a positive claim than a similar, but negative one, links well to the second reason that I want to suggest for why Dretske does not offer systems that are capable of second level intentionality, it is this; the stratification he attempts to make is intended to differentiate between mental states of different kinds and this is a very difficult thing to do since mental states are not discrete, discontinuous entities. Indeed a lot of the confusion that is encountered in chapter seven of *Knowledge and the Flow of Information* is a direct result of the difficulties that are bound to be encountered when anyone attempts to force vague or fuzzy[15] concepts into an explanatory structure that suits only those sets of things that are limited and distinct.

Mental states cannot be measured like machine states can

In the Chomsky hierarchy we saw that he was dealing with machine states that are finite, limited, and thus measurable. The Dretske hierarchy on the other hand deals with quite different entities, mental states which are not finite and measurable in the way that machine states are. In the case of a machine state the static state can be quantified, the input quantified and the output, whether determinate or indeterminate, also quantified. Everything about the machine and its states is limited by its architecture and the function for which it has been designed. Its functions are already known or knowable.

Measurements can be quantified because they are discrete chunks of information, but mental states are not in such a specific form and it is for this reason also that I think Dretske has failed to suggest any system that is capable of knowing but not believing. The difference between being a simple processor of information and being capable of doing something more with the information is plainly seen, but trying to calibrate the

different stages of mentality that occur after the information processing is probably not even possible, except in a very primitive sense.

Again the problems of his peculiar usage of analogue and digital create difficulties for in its more common usage digital pieces of information are those discrete pieces that can be calibrated, whilst analogue information is in a continuous and unquantifiable form. In Dretske's explanation of perceived information the signal is analogue in form, which is fine because it is vast and unmeasurable, however the specific semantic content of that signal is said to be in digital form and thus a discrete piece of quantifiable information; but Dretske equates being able to reach the semantic content of a piece of information with third level intentionality and the ability to form beliefs about a world and, as has been argued, what something means and the mental state of holding a belief are not measurable things.

It is not possible for me to say with complete conviction that, "My cat knows there is food in that cupboard but it does not believe that there is food there", for it is not possible for me to draw a hard and fast distinction between those two mental states in any other system. Even in my own case it is not easy for I can say about myself that "I know that there are people who believe and worship a God of some kind, but I do not believe in the existence of God", but I cannot say "I know there are people who worship a God, but I do not believe that there are people who worship a God" for it does not make sense.

I distinguish between the different mental states of other human beings on the basis of what they tell me using propositional attitude statements and through watching their subsequent behaviour. If they pray, attend religious gatherings and observe religious festivals then I can conclude that they believe in God. Nevertheless the difference between knowledge and belief states is still a very difficult one to set out, for, as we have seen briefly, what can be the difference between my knowing that something is the case and my believing that it is the case. In the *Philosophical Investigations* [16]

Wittgenstein says that when you are certain of something is it not just "shutting your eyes in face of doubt?"; and when you are certain of something you feel you know it for sure, that is, beyond doubt. For the mental state of 'belief' it must be much the same since a belief in God is really only a "leap of faith". It is, after all, simply a decision to stop doubting.

Another problem with knowing and believing is that to assert that I know something and at the same time deny that I believe in the same thing is to speak nonsense, but to assert a belief in something yet deny any sound knowledge of that thing is indeed sometimes very sensible. For example, I can say that "I believe with the particular cloud formation that it might rain", but I cannot say that "I know it will rain", nevertheless my belief might prompt me to carry an umbrella with me when I go out thus stacking the odds against me getting wet whatever happens.

The problem of distinguishing between a knowledge state and a belief state is just as difficult for concrete examples such as bus timetables or thinking that a particular public house serves the type of cider you like best. For example, I believe that on most of the occasions I have been able to catch a bus from the city centre to home at either ten minutes to the hour or twenty minutes past the hour, am I now in the position to say that I know a bus will come at these times? I think not, for often I have waited and the bus has not arrived. On each occasion that it has not arrived I have repeatedly gone to the timetable to check that the information I have is correct and that one is due, but although I 'believe' that a bus is due I would never say that I 'know' that a bus will come. The same goes for the example of the public house, for although I 'believe' they stock the brand of cider I like I also know that some times when I have gone they have not had a delivery and this means that I can never say I 'know' that they will have that cider when I go this evening.

Belief and knowledge states of my own are problematic enough, but when I try to distinguish between what counts as a belief and what counts as knowledge in another

system, and worst of all a non-human system, I am doomed since our concepts for mental states are, as yet, hopelessly vague. The representations that organic systems have of their worlds are only ever approximations of reality which can never be entirely accurate for they are the function of all sorts of perceptual limitations. For instance, the fly cannot see the open window above it through which it can fly to freedom because it is limited by its perceptual apparatus. It is not possible for any system to reach the 'perceptual phase' or 'finite point' of all knowledge, that is, for the system to know all that there is to know, because finite knowledge represents an ideal state of knowing, of knowing everything that is knowable of which only the omniscience of a god would be imagined capable.

It is easier to see the inaccuracy of an arithmetical approximation than to see the inaccuracy of our semantics that are based on the knowledge we possess or the limits of all our possible knowledge. With arithmetical and mathematical models we measure things that can be broken down into discrete chunks, and machine states are of this sort. In *Events and Reification* Quine proposes some individuating criteria that hold for physical events, "Physical objects are well individuated, being identical if and only if spatiotemporally coextensive". Mental states do not fulfil such spatial or temporal criteria and we cannot measure arithmetically those (mental) states that are continuous and only vaguely distinguishable from one another. And, in *The Individuation of Events* Davidson says of the individuation of mental events or states that "We classify an event as mental 'if and only if it has a mental description, or ... if there is a mental open sentence true of that event alone'. An 'open sentence "event x is M" is a **mental** open sentence if the expression that replaces "M" contains at least one mental verb essentially.'" Chomsky's hierarchy does not deal with events of states for which a mental description is possible, rather his hierarchy is one of absolutes where a straightforward set of machine states and tasks can be described and set out in a limited number of discrete steps.

On the other hand the business that Dretske attempts, of distinguishing between mental states, is not one of absolutes but one that is best attempted using the strategy that taxonomists have adopted, which is to map clusters of points according to their similarities or overlapping characteristics. In this way it has been possible to say of something, say a slow worm, that it carries characteristics of both lizards and snakes, for it has relics of shoulder and hip bones that mean that it was once an animal with legs such as an iguana or a skink, but now these legs have proved redundant and it has become more like a snake. Thus from its characteristics taxonomists can confidently place the slow worm in the species: reptile.

At the beginning of chapter six I will make some suggestions for ways of dealing with vague concepts that are better than the present attempt which has been to stratify and form a hierarchy of them. For now I will look at the fourth, and final, criticism I will make of Dretske's division of intentionality and, more importantly, mental states and capabilities.

5.4.4. Genuine cognitive systems

Dretske distinguishes between systems that are simple conduits of information capable of only a first level of intentionality and 'genuine cognitive systems' that are capable of third level intentionality and also of first and second level intentionality as well. In this second category of systems he places 'frogs, humans and perhaps some computers', presumably all of which are capable of forming beliefs about their worlds. But I find his phrase confusing, for what exactly does he mean by a 'genuine cognitive system' and why does he include inorganic systems in this category, when Rosenschein goes as far as saying that systems such as these are logically only capable of a second level of intentionality, that is, having epistemic states.

In the category of information processors there are no organic systems of any type, only inorganic systems, televisions, thermostats and so on. These have a simple

function, they react to a particular type of information signal in a particular type of way. Their activities are entirely dictated by their design. They can do no more for they have no flexibility to adapt to their environment, and can only do less if they are damaged in some way or without any power supply.

The other category of those things that are genuinely cognitive is an altogether more interesting one for there are systems in it that are made of silicon alongside those that are carbon based, for Dretske seems to be distinguishing, not between mental states and machine states, but between systems that process information and systems that form beliefs. Those that can form beliefs are also those that are able to select appropriate information whilst ignoring other unimportant pieces, storing other information for later use, understanding the information that has been selected, analysing it, conceptualising it and forming beliefs about it that will change the patterns of other beliefs that are held or form the basis of new belief structures. These are capabilities of which the simple information conduit is not capable.

It is this issue of 'belief' that is significant for Dretske in the formation of a distinction between information processors and genuine cognitive systems. Being able to form and hold beliefs is a necessary characteristic of any genuine cognitive system and something of which simple information processors are not capable. I shall not argue with this for the moment; first I shall explain his position in relation to the 'information-theoretic' account.

In his example he states that the curvature of the bimetallic strip inside an ordinary home thermostat is what registers any change in the temperature of the room. The degree of its curvature regulates the heat by touching a contact in the adjustable heat control in the room. The thermostat is dependent upon the strip which according to the information-theoretic account is a rather primitive heat detector. Thus for Dretske, "A belief is like the configuration of a bi-metal strip in a properly functioning thermostat: it is an internal state that not only represents its surroundings but functions as a

determinant of the system's response to those surroundings".[17] The second half of this sentence is of most interest for it proposes that a belief is not only the result of interaction between a system and its surroundings but also that it is the cause or determinant of the systems subsequent behaviour. It is safe to argue from their status as 'simple information processors' that Dretske does not wish to claim, as McCarthy has before him, that thermostats have beliefs, for as he says in the footnotes of chapter eight, their "internal states have no appropriate semantic content".[18] For the thermostat the curvature of the strip has no meaning yet it does determine its future action.

So for a system to show that it is capable of forming and holding beliefs there are three signs that it has to exhibit; firstly, there has to be a loss of information between its perceptual input and its conceptualisation, secondly, the beliefs have to be related to the system's environment by being formed as a result of it, and thirdly, the beliefs have to determine subsequent behaviour.

Human beings certainly exhibit all three of these characteristics but I am not so sure about other systems. For instance, it is certainly the case that cats and frogs ignore a great deal of the perceptual input from their environments, and that they select only that information from their environment that is of relevance to them; but surely this can be said of ants and flies for they too only seem to respond to the things that are of immediate relevance to them. But this might simply be because the ants and flies only possess the perceptual apparatus to respond to a very small part of, what for us is, a very large world. Cats, frogs, horses, weasels, and so on have a perceptual apparatus not unlike the human one so their world is more likely to be on a par with ours because of this. But it might also be because of their size, for we can see them react to things that we too can perceive. For example, I can see my cat's ears twitch when it hears me opening a can of baked beans which it has mistaken for a tin of cat food, but I am unable to see the movements of the fly's eyes when it watches me coming closer with a newspaper to swat it. The noise of a tin being opened certainly does make my cat

consistently behave as though it believes it is going to be fed but I am more inclined to describe this as learnt or conditioned behaviour rather than a belief that the cat possesses and stores up for future use to determine its behaviour.

It does not seem possible to say of non-human organic systems that they are 'genuine cognitive systems' in the sense that they can form beliefs and use those beliefs in the way that human beings form and utilise beliefs. Yet it does seem possible to say that non-human organic systems are genuinely cognitive in the information-theoretic sense, set out by Dretske, for then it means only that they have internal states that represent their surroundings and also serve to help determine their future responses to their surroundings.

But Dretske includes not only organic, but also inorganic systems ('perhaps') in the class of that which is genuinely cognitive. This can be accepted but again in an information-theoretic sense for such a sense is heavily constrained by what the system has to be capable of doing. That is, in the information-theoretic sense the system does not have to be capable of as much as it would in the natural or realistic sense of what would count as genuinely cognitive. For instance, it would be hard to accept computers as 'genuinely cognitive' in any but an information-theoretic sense for such systems have finite, measurable states, they cannot perceive anything beyond their pre-programmed domain, they cannot form beliefs as the result of analysing and understanding the stimuli to which they have responded in their environment, nor can they offer any subjective interpretation of the information they perceive, and finally, mitigating against all non-human systems, organic an inorganic, is that neither type of system can form beliefs about abstract concepts in the way that human beings can. No computer, except perhaps those in the realm of science fiction, for example, "Hal" in the film *2001*, can ruminate over the mysteries of life, the problem of identity or the existence of God.

5.5. In conclusion

It seems that the difficulty about what is, and what is not, genuinely cognitive is a problem that results from Dretske's not being able to offer a possible system that is capable of knowing but not believing. It is not possible to say precisely what counts as proof of cognition for the elements of cognition are mental states and they are not defined in the way that machine states can be. It is doubtful that 'some computers' are genuinely cognitive for the internal states and structure of the systems that we consider to be cognitive in any sophisticated way are quite, quite different. That they might be artificially cognitive is something that is already accepted for machines can be designed to behave 'as-though' they have a particular type of mental state that equates with cognition of a specific kind.

Much of the problem about what is, and what is not, genuine cognition is also part of the long running problem about what counts as one type of mental state whilst not counting as another. That is to say, when does my 'liking' turn to 'loving', my 'hopes' to 'desires', my 'knowing' to 'believing', and so on. The distinctions between one sort of mental state and another, or even the different levels of intensity of one particular mental state are difficult, if not ultimately impossible, to draw up. To attempt to distinguish between the mental states of different human beings is a vast task that has all the advantages of analogies between behaviours and a shared, descriptive language. To extend this distinction to look at the mentality of different organic systems is yet more complex for all we have to go on is the other system's behaviour since there can be no shared language. To move another step further and try to look for similarities and distinctions between mental states and machine states is yet more difficult because there is only the machine's already programmed behaviour from which we can draw any comparison, and this behaviour is itself the product of human creation. A sort of 'homo

ex machina', which might suggest that machine actions and states be thought of indirectly as second-hand human actions and thoughts.

Every organism, human or non-human, has a mental life that differs from every other organism in form and content. For example, as a human being the content of my mental life is distinct from the mental lives of any other human beings, but the way I process, store and use information, that is, its form, is something that I have in common with all other human beings. It is likely then that this is also much the same for any commonality that I have with higher order animals. So that the form of my mental life and constituent mental states will be much the same as the form of the mental life and states of an orang utan or a chimpanzee, but that this commonality becomes less and less so as I compare my mental life with animals lower down the phylogenetic scale. Thus when I reach a comparison between the my own mental states and the machine states that accompany the computation of a machine there is very little similarity to be drawn, but still there is some and this will be a matter that I shall attend to in chapter six.

In chapter six I shall show that this commonality, or perhaps the significant lack of it, can be better shown in cluster diagrams than in stratifications and hierarchical arrangements. As mentioned earlier, diagrams of this sort are often used as taxonomic devices for deciding the category of one species or another. I will be using them in this context to express the overlapping nature of mental states and in which systems such states can be said to exist in some form or other. In this manner I will also show that there is some overlap between the capabilities that I possess as a complex human organism and the capabilities that a thermostat possesses, and that the thermostat is much more efficient and capable at what it is designed to do than I would ever be because my design and functionality is necessarily different from its.

Endnotes:

[1] The section on Chomsky is taken from Krishnamurthy, E.V. (1983) *Introductory Theory of Computer Science*, Published by Macmillan Computer Science Series.

[2] A *transition graph* resembles a flowchart consisting of labelled circles that represent states and arrowed or directed lines that either loop or go on to another state or circle. The input state is indicated by an input arrow and the final state by two concentric rings.

[3] Krishnamurthy, E.V. (1983) *Introductory Theory of Computer Science*, section 5.6 ff.; Macmillan Computer Science Series

[4] Although I use a sentence of natural language to exemplify derivation this grammar is still very limited and can only be used to generate very simple, unambiguous sentences in natural language. It is still more suitable for generating propositions in a formal language.

[5] A full account of this hierarchy is set out by Dretske in chapter seven of *Knowledge and the Flow of Information*, (1981) Basil Blackwell.

[6] By the 'same sensory input' I mean reading the same article in a newspaper or looking at the same painting in an art gallery. I do not mean that they could ever have the same perceptions that would be identical from every angle and with the same personal history, for this would have to mean that they were the same person which is logically impossible. Kant's theory of 'Incongruous Counterparts', (*Critique of Pure Reason*, 1787 Macmillan (1929)) gives credence to this view, but I'm sure that some of the contemporary studies that concentrate on twins and multiple births might suggest that two or more people that are born together can have perceptions that are essentially the same.

[7] Chapter two, section 2.8.4. "Example of focusing and selectivity"

[8] By the 'extent of its domain' I mean here the amount and variety of interaction that any system has within its own environment.

[9] Dretske, F (1981) *Knowledge and the Flow of Information*, chapter 7, p.177, Basil Blackwell

[10] I am not claiming here that the analytic/synthetic distinction is tenable; I am only using it, as Dretske does, as a suitable descriptive term.

[11] Dretske, F. personal communication (Email)

[12] As we have seen in chapter two Rosenschein does this when he assigns primitive epistemic properties to machines, but only those that can encode their knowledge in an appropriate formal language.

[13] Gettier, E.L. (1963) *Is justified true belief knowledge?*, Analysis 23.6, p.121-123; and also cited in four other places in Gettier's footnotes as "*Theaetetus 201*, and perhaps....*Meno 98*", "Roderick M. Chisholm, *Perceiving: a Philosophical Study*, Cornell University Press (Ithaca, New York, 1957), p. 16." and "A. J. Ayer, *The Problem of Knowledge*, Macmillan (London, 1956), p.34.".

[14] This is similar in kind to Wittgenstein's and Malcolm's argument against analogy for thinking that another person has mental states like mine, for they would say that I do not go through the process of thinking to myself, '*x* is crying, and they resemble me bodily, and every time I cry I am unhappy, so *x* must also be unhappy', but I would argue that the process is there nevertheless and that it is something we grow up doing and learn to do implicitly, that is, without it being accompanied by a linguistic affirmation.

[15] In this context I use 'fuzzy' to describe concepts that cannot be delineated from other concepts of the same kind, and the concepts 'of the same kind' are 'mental states'. No reference, overtly or otherwise, is being made to the area of fuzzy logics.

[16] Wittgenstein, L (1958) *Philosophical Investigations*, Section II xi, p.224, Basil Blackwell

[17] Dretske, F (1981) *Knowledge and the Flow of Information*, chapter 7, p.198, Basil Blackwell

[18] Ibid. chp.8, p.261-262 (footnote 6)

6. Illustrating vague concepts

6.1. Introduction

This chapter will have the following structure. As in the last three chapters I will begin with a statement of the problem area and then take a look at the specific question, or as is more relevant in this chapter, the particular issue that is to be confronted. I will then begin the main body of the chapter with a reiteration of the main conclusions so far and explain why these relate to the necessity for a more successful way of demonstrating the correlation between a system's internal states, whether mental or machine, and that system's complexity of architecture. Following this I will move on to give examples of some of the alternative ways in which the relationship can be illustrated and that each of these, although limited in their own ways, is still better than the attempts to produce stratified hierarchies. I will attempt to show that no perfect set of axes exists within which the nature of vague concepts can ever hope to be adequately defined, from which I can only but conclude that it will never be possible to describe mental states in absolute terms.

In the next part of the chapter I will look more closely at the recent work of Aaron Sloman for his work concentrates on design and the 'design space' in which different architectures occupy different points. Sloman argues that for a system to be capable of different activities it would need to occupy different points in the design space. Thus for a system to be capable of more complex things it needs a more complex design space. For Sloman the human being has a very rich and complex design space and it can be inferred from this that it also has a rich and complex repertoire of possible behaviours. But being rich and complex is not sufficient for the performance of complex behaviour for in addition to the design space we need also to look at what the system needs to sustain its existence in the environment it occupies. In other words, what it is that keeps the system alive and functioning.

Three main conclusions will be derived, and along with these will be a number of lesser conclusions concerning the requirements of a system for it to be capable of exhibiting the behaviours that it does. The first of the main three conclusions will be that every graphical interpretation of a state of affairs is unique since it will always depend upon what is being examined, or 'plotted' on a cluster diagram, and what the things being examined are to be measured against. If I were to choose a different set of axes many of the systems would not appear at all and we would perhaps be looking at more specific information about fewer systems. This is one area in which further work could be carried out. The second main conclusion will be that two dimensional representations are inherently limited and an increase in dimensions, and as a result accuracy, is absolutely necessary if it is going to be possible for us to establish any relationship between the mental states of a system, its complexity of architecture, its ability to adapt and its overall behavioural capabilities. There are just too many criteria. The third, and final of the main conclusions will be that by using the taxonomic method of description or display it has at last been possible to show a comparison between the differing capabilities of a wide range of systems, whether organic or inorganic; on the basis of this comparison an examination of machine states and mental states using the same criteria for each will have been made possible. This has distinct advantages over the hierarchical stratifications of mental and machine states that have been favoured by people such as Dretske in previous work. From the taxonomic representation of information it will be possible to deduce how likely it is that different types of mental states are present within systems that are essentially quite different from human beings.

In the final stages of the chapter I will concentrate on some of the other conclusions, with references being made not only to what a system does, but also to how it is capable of carrying out such actions. Which is to say, what mental states, other than straight-forward adaptability are required by a system for it to be capable of those actions that it needs for its continued survival.

6.1.1. A statement of the problem area

Mental states are fuzzy in the sense that there can be no clear delineation of where one state stops and another state starts. They have no clear-cut beginning or end in space or time and for this reason it seems implausible to stratify mental states, setting them out in some sort of definitive hierarchical model. In chapter five I have argued against such ways of envisaging a relationship between a continuous set of mental states, and I have argued for the acceptance of such models for exemplifying machine states which can be differentiated.

In this chapter I will argue that it is still possible to draw a comparison between mental states and machine states, and although the area of commonality is always shifting and changing with the influx of more information and the creation of new and more complex machines, it remains a relationship that can be shown using a taxonomic device. Indeed it is certainly the case that using a taxonomic device, such as a cluster diagram, is beneficial because it permits a point that represents a particular type of machine to be shifted if that machine is, for example, redesigned to possess new capabilities or an increased domain. This is just the same for a living system that might adapt to its changing environment and develop, over some lengthy period of time, a different or improved capability. A second advantage is that the shifting of a single point does not affect any of the other representational points in the diagram, and nor does it call for the redefinition of other points or axes in relation to the changed status of the one altered point.

6.2. What has brought us to this stage?

In this section I will briefly recall what has been said from chapters three to five to give some indication of how we have reached this stage. An analogy with the text can be seen in the art of weaving for the cloth can only be kept in good shape if the threads remain taut and even. This section brings the 'threads' of the argument together so that they are kept 'taut' and the pattern of argument can be seen to emerge.

6.2.1. Chapter three - ascription

The notion of ascription was examined in chapter three and in particular how a human being actually sets about ascribing mental states to systems other than itself. A number of important points were brought together in the conclusion. The first was that ascription is usually made on the basis of at least two criteria, (i) that the behaviour of the other system is consistently human-like so that an analogy can be drawn with one's own behaviour, and (ii) that our apprehension of the other system's architectural complexity is such that we might think it feasible for it to have mental states. There is a third factor that is influential, but only to our ascription of mental states to other human beings, and it is that we share with them a language, through which we can proffer confirmation or denial of any ascribed state.

Thus the ascription of mental states is by no means simple for all of the criteria depend upon our subjective view of our world and the information we receive from it. For example, I might say of Rose that "She knows what she is talking about", whereas you might think she is deceiving us rather cleverly, and neither of us would be wrong in any strong sense for our opinions of Rose are based on our own personal, and ultimately subjective, points of view. There is no decisive view to have, for Rose might truthfully believe that she knows something when in fact she has only been lucky not to have been asked difficult questions, or you might know a lot more about the subject and feel that what she knows is only a paltry amount, or you might know nothing about the subject and feel envious of her knowledge. There are a great many possibilities when dealing with the mental states of another being and our own subjective judgements. That Rose might have one of any number of mental states, none of which can be pin-pointed with any high degree of accuracy, give us some indication of how confusing is the business of mental state ascription. Inevitably then, our claim must be that if another system does possess mental states only it can ever know for sure that it has them.[1]

6.2.2. Chapter four - complexity

The original question of chapter four was "Given a specific task or competence, what is the minimum system that would be required to accomplish it ?". My enquiry began by looking at three approaches to the notion of complexity in relation to living and non-living systems, (i) the architectural complexity of the system, (ii) the complexity of the system's actions or behaviour, and (iii) the complexity of the relationship between the system and its environment.

I concluded the first section by stating that a marked relationship could be seen to exist between the overall complexity of a system and its capabilities to perform certain actions. In just such a way then the capabilities of a computer are dictated by the combination of its architecture, the program that has been instantiated and the environment in which it is fixed. Machines of this sort have no flexibility to choose what information they will react to in their environment for it is all part of their pre-programmed design.

Similarly the capabilities of non-human animals are also widely dictated by their environment, but when we look at higher-order animals we discover that they have the added capability of being able to choose what they will attend to in their environment. Therefore they have the added advantage of being adaptable. From this selection they can choose how they will respond to the information, for example they might wish to run away, to conceal themselves or to fight. Such a response as this might be the result of a specific genetic structure, and in some sense 'pre-programmed', but with animals such as monkeys and even cats, the possibility of a self-consciousness element to their judgements cannot be ruled out completely.

When dealing with human beings it is possible to say, but only with reference to my own experience, what they can and cannot do. I know that I am capable of processing vast amounts of information, selecting what are the most important pieces for me, responding to them and storing for later use what is not immediately required. And what is more, I can do all of this with myself at the heart of my judgements. I

interpret all the information I receive subjectively so that any information I pass on to other people will have the addition of my own point of view with those pieces of information left out which I feel are irrelevant. As a product of my environment I can act in my own best interests, but as the self-conscious product of society I am also capable of subjugating my own interests in the interest of the continued survival of society as a whole.

To act in its environment any system has to be capable of processing information and this requires a certain amount of awareness. Such awareness is exhibited by all systems, from the most limited to the most flexible, by their capability to react to stimuli that are relevant to it. But for a thermostat to respond to a rise in temperature indicates only that the system has a very limited range of actions and no flexibility to choose between relevant and irrelevant stimuli at all. So a simple awareness only shows that the system, like a thermostat, can respond to those aspects of its limited environment for which it has been designed or programmed.

Thus the complexity of a system was seen to relate, not only to the internal and external architecture of the system, but also to the degree of flexibility that the system has to respond to a variety and changeable number of stimuli within its environment. In the human system a better way of describing this 'flexibility' might be to say 'versatility', for 'versatility' is usually associated with the idea of 'being able to turn one's hand to anything' and the human system is indeed capable of responding to a tremendous wealth of informational stimuli.

As a human system I am capable of many things, high-level awareness, the selection of relevant information, understanding that information and making self-conscious judgements involving it. I am also able to anticipate, to some degree, how other objects and states of affairs in my environment will be affected by my judgements, and to change my judgements, try to justify them to others or try to change the judgements of other people. I, and all other human beings, if my extrapolation from myself as an example of human sentience and experience is truly valid, are very complex systems indeed with a great many capabilities.

6.2.3. Chapter five - stratifications and hierarchies

Chapter five began with a look at a Chomsky's hierarchical stratification of machines and their respective capabilities to recognise and interpret grammars of varying complexity. The sorts of capabilities that a machine can exhibit depend very much upon its internal states or architecture and its auxiliary memory that stores and retrieves information. Thus a machine as basic as a Finite State Machine (FSM) has only a very limited set of capabilities, whereas an unbounded Turing Machine (TM) is capable of almost anything theoretically.[2] A relationship can be seen to exist between the structure of a system, its link to its environment, or domain, and the things of which it is capable. The tasks are described as 'certain' because Chomsky defines them for us and they consist only of the recognition and interpretation of four different types of phrase structure grammar.

Chomsky does not deal with mental states so Dretske's stratification of intentionality, the mental states that correspond to the levels and the systems that are capable of achieving each level, was examined. It is an 'information-theoretic' account that examines which level of intentionality the system exhibits based on the quantity and extent to which it can process information. Dretske attempts to show that simple systems, that are capable of only simple information processing, can occupy first level intentionality and only the most basic of mental states. More complicated systems that can exhibit some understanding of their incoming information are correspondingly capable of a higher level intentionality and, therefore, also of possessing higher level mental states.

Many difficulties were encountered with Dretske's stratification. The main ones were that he finds it impossible to offer any system that can be said to 'know' yet not to 'believe', and that he gives only a faint idea of what is meant by his phrase 'genuine cognitive system'. I have argued that these two problems are inextricably linked because of the problematic nature of saying precisely what factors go to make up a mental state. For example, wherein is the difference between 'knowing' and

'believing'. My only certainty when it comes to mental states is that I have them, and by analogy it is feasible for me to conclude that other like systems also have them. It is not as feasible for me to presume that other non-like systems have mental states, so what I base my ascription on then is their behaviour being human-like.

This problem with defining and differentiating between mental states is such that it makes the relationship between capabilities and complexity easier to observe in machines for their states are fixed and measurable. Mental states are vague and unmeasurable, with indistinct boundaries where one overlaps with another making the discernment of a single type of mental state nigh on impossible.

If we accept that all living organisms have a mental life of some degree no matter how limited, each different species interprets its world in its own unique way, so that the form of the mental life of any one species will be different from the form of the mental life of any other species. Thus there is a difference in how each species receives and processes information and it is this that makes it possible for me to say of another human being that she 'knows', 'wishes', 'hopes' or whatever, but not so likely of any other species that they have mental states that are identical to mine.[3] However, when it comes to the content of each systems mental life it is something which is unique to each member of each species, for I can never have the experiences of another human being, or for that matter, another species. It is the content of my mental life that makes me distinct from all other human beings but its form that unites me with them.

A commonality or overlapping exists between my mental states and those of the other higher order primates , but it is a commonality that lessens as we descend the phylogenetic scale. When I go as far as to compare my mental states with those of a machine very little similarity can be drawn, but what there is increases as the machine becomes more capable and is able to perform tasks of which I had thought only myself and other human systems capable. Indeed there are many tasks that a machine is distinctly better at performing than a human system, and it is just this type of anomaly that Dretske's stratification fails to show. I shall now attempt to rectify this.

6.3. The relationship between vague concepts can be shown

Throughout this thesis it has become clear that both the mental states that we can be said to possess or be 'in', as for instance "I am in a state of shock", and our concepts of what mental states are, that is, how we define them and the contexts in which we use them, are altogether muddied and vague. Beginning with the notion of their recognition and ascription and right up to the problems of differentiating between them the area is consistently beset with problems. These problems are such that they diminish the real possibility of any coherent study being carried out. This being the case I would like to show that although the relationship between vague concepts is not clear cut like the division between machine states, all is not lost, for they can still be illustrated and discussed using cluster diagrams. Diagrams of this kind are capable of showing where any overlapping concepts or states are most likely to occur, and of thus creating a way of viewing mental states in a fuller context. What I mean by 'fuller context' here is that the concepts or states are placed in relation to others of a similar kind with which they might not usually be seen to bear any direct relation. But again it must be stressed that any relationship that is established will be constrained by the axes that we choose to use. Thus it is likely that were we to choose a different set of axes the same groupings or clusterings would not show up.

In the section that follows I will offer some examples of cluster diagrams that set out to show the relationships that exist between mental states and the systems that can be said to occupy them. Provisionally the two axes that I will use are architectural complexity and the limitation on the system's flexibility to behave, in broader terms, their range of capabilities. I will discuss the extent to which each diagram manages to fulfil the purpose for which it has been drawn up, how the representation might be improved and what conclusions can be drawn from attempting to exemplify vague concepts in this particular manner.

6.3.1. A bit more about the concept of cluster diagrams

Cluster diagrams are usually two dimensional representations that are used to indicate the 'clustering' of objects or entities into groups on the basis of their bearing like characteristics. It is a strategy adopted by typologists and taxonomists that allows them to classify animals, for example, into a particular species or to show a succession between one type of species and another. Diagrams of this sort show a continuity between species that is not possible to see when that species is viewed as a section of a stratification or hierarchical arrangement. The procedure is that individual types of entity are marked at points along a pair of axes and it is argued that those that fall most closely together are related on a basis of some overlapping characteristic or characteristics. In taxonomy these are most likely to become members of the same class or species, and where their is greatest diversity it is a very useful technique.

A good example that shows the necessity for this method of classification is among the beetle family, where the weevil group alone has over 40,000 different species. The weevils are grouped together on the basis of a 'rostrum' or protruding snout. Half way along the rostrum is a pair of antennae and at its end is a set of jaws. Most weevils are flightless, scaly and have a vegetarian diet. Their larvae are usually legless and feed and develop inside the food plants in which they have been laid. So these are the characteristics used by taxonomists to categorise beetles of the weevil type. Thus any beetle which possesses these characteristics, or at least the majority of them, will be categorised as a member of the weevil family.

The idea of looking for the features that one thing shares with another bears a great similarity to Wittgenstein's notion of 'family resemblances'. Wittgenstein uses the example of "games" and he argues that there is no one game, of which a thorough understanding would tell us that that is what it is to be a game.Which is to say there is no one single feature that all games have in common that we could say that any time the feature arose then what we would be playing or partaking in was a game. The concept of 'game' is only something that can be grasped by looking at lots of different games

and types of games and trying to pin-point their overlapping characteristics. So games such as Bridge or Backgammon have an element of competitiveness in common, whereas the game of "ring-a-ring-a-roses" is only played to have fun and no competition exists between the players. On top of this we also talk of the 'games' that people 'play' in relationships which further complicates the issue of what it is to be a game. The question we should ask is what are the elements of this kind of interaction that permit us to describe something as a game; our answer might be that the shared elements are those of enthusiasm, enjoyment or a desire to win. "One might say that the concept 'game' is a concept with blurred edges."[4]

Wittgenstein goes on to liken this to what happens when we look at some of the members of a family group and what allows us to recognise them as members of the same family. For instance, what makes it possible for us to say of Ian, the son of Jane and Barry, that he has 'his father's nose' or 'his mother's eyes'. The answer is that what we look for and pick out are the 'family resemblances' that exist between 'members of a family'.[5] Such resemblances are the commonalities of feature that somehow manage to bridge the uniqueness of every individuals DNA structure and the gap between different generations of one family making it possible for Ian to be recognised as 'his father's son'. They are not features that every member of every family shares for then they would be clones with no differentiation between them. So Ian's sister may also have her father's nose but have her grandfather's eyes and her mother's smile. In this way they can be differentiated whilst still being recognised as members of the same family group.

The same sort of family resemblances between objects or entities can be seen in the groupings of plotted points on a cluster diagram. Each diagram that will be shown will be followed up with a discussion of its merits and demerits.

In my diagrams different systems will be clustered together on a basis of their complexity of architecture and capabilities. These, it has already been said, are its constraints from which we can only ever obtain a limited picture of the true relationship between systems, but then in a two dimensional representation all things cannot be

considered. I think I should emphasise at this point that these diagrams are intended only to give some idea of how a two dimensional cluster diagram can be represented so none of them should be considered as wholly accurate or final.

One of the main conclusions of this section will be that a two dimensional representation is too limited for what anyone requiring a concept of the mind could hope for, and that it would surely be better to look at the relationships using a three, or perhaps even four, dimensional diagram. Diagrams of this sort would themselves be limited but their big advantage is that they could contain a lot more information than we can now envisage on a two dimensional framework, but a lot more about will be said about this in section, 6.3.2..

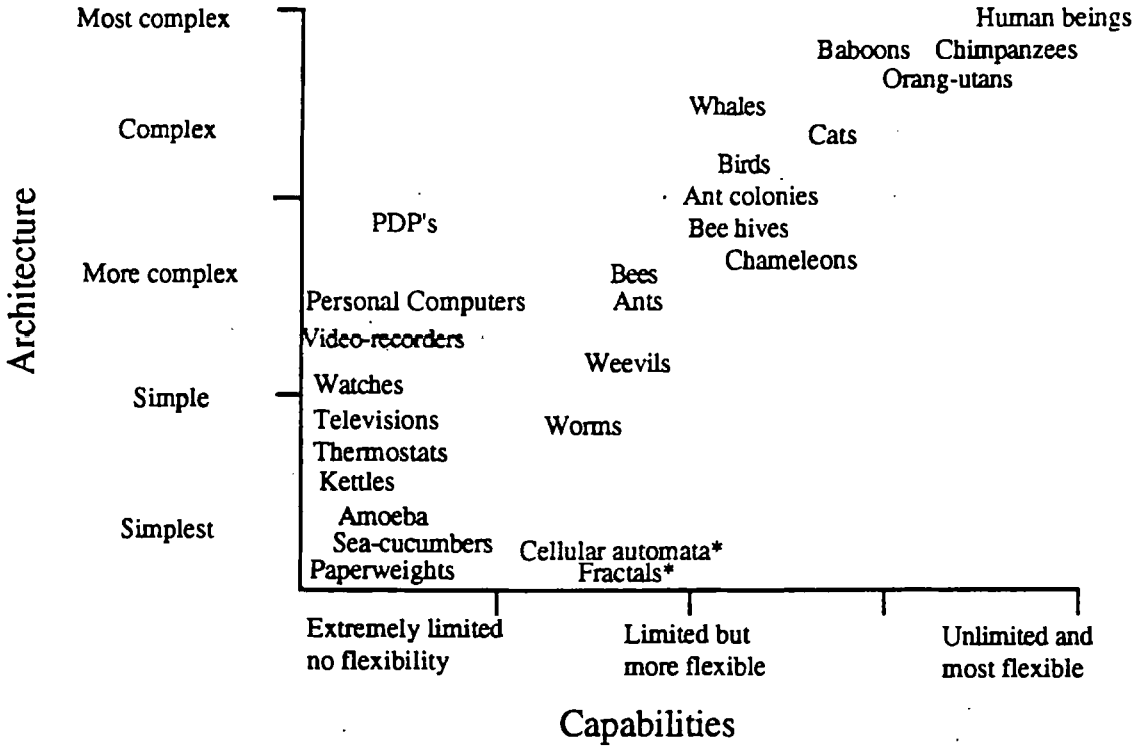

Figure 14

In this first diagram I have plotted the architecture or internal structure of a small number of systems alongside their flexibility to exhibit particular capabilities or sets of capabilities. At this stage the two axes have been left deliberately general but even so it is possible to see that a strong correlation exists between the two. The first thing to notice is that there are at least two rather general trends that have emerged. The first is

that the inorganic entities, such as kettles, televisions, watches, and PDP's have become gradually more complex in design but only very slightly more capable, thus they occupy the lower to middle left-hand side of the diagram. A machine's flexibility is limited by the complexity of its design and the overall function for which it has been designed. So, as I mentioned in an earlier example, a thermostat cannot make tea nor groom the dog because it has not been designed to carry out these functions. Had it been necessary for it to possess these capabilities, as well as being able to detect subtle changes in temperature, it would have had a more complex architecture and internal control mechanism. However, it would no longer be an example of a commonplace thermostat.

The second trend can be seen among the organic systems for they tend to move in a fairly continuous and non-arbitrary fashion from the lower left-hand corner to the upper right-hand corner. There is very little deviation from the central diagonal line which suggests that the respective complexity of each system is very firmly linked to its capacity to act with different degrees of flexibility to those systems that have different architectures. There is nothing in the bottom right-hand corner of the diagram nor half way along the bottom, but this mid-way point would be reserved for systems that have elaborate capabilities but very simple architectures. It is unlikely that this section of the diagram would ever have many occupants but there are some and they are those that are marked with an asterisk, the 'Fractals' and 'Cellular automata'.[6] These are exceptions to the 'rule' for they share a special status, and it is this; they are each simple systems that have a limited flexibility but are yet capable of a great deal of complex activity.

The Julia or Mandelbrot Sets are very simple fractal equations that can produce complex, recurrent patterns. Some of the more common examples of this sort of complexity can be seen in the edible flower of the cauliflower or on the fronds of any of the family of ferns, such as *Pteridophyta*. A further example of a simple system being capable of immense complexity can be seen in the *Fibonacci* series of numbers. A series of numbers where the consequent is always the sum of the pair that precedes it.

A natural example of this series can be seen on the coarse skin of the pineapple which descends and ascends in a spiral pattern.

Likewise, the "Game of Life" is an example of a simple cellular automaton that is capable of manifesting a great deal of complexity. The principles of the game are such that if one square, representing a cellular automaton, has no other square beside it nothing happens and it remains stable. If there is one other beside it the first square dies. If there are two squares beside it they both stay alive. If there are three around the first square then another square is formed and finally if one square is surrounded by four other squares the central one dies. Thus some squares are brought into life whilst others die by colliding with others or by being collided into. There are some arrangements of squares that have, at least, a temporary stability; these are the single square, the lozenge shape, four squares together forming a single bigger square, and any number vertically or horizontally arranged in a line.

The following diagram shows a only small selection of the arrangements of squares that can be produced in the "Game of Life".[7] However, it is still possible to see both the simplicity of the cellular automaton and the complexity of a few of the hundreds of possible arrangements of cellular automata from what is an essentially limited diagram.

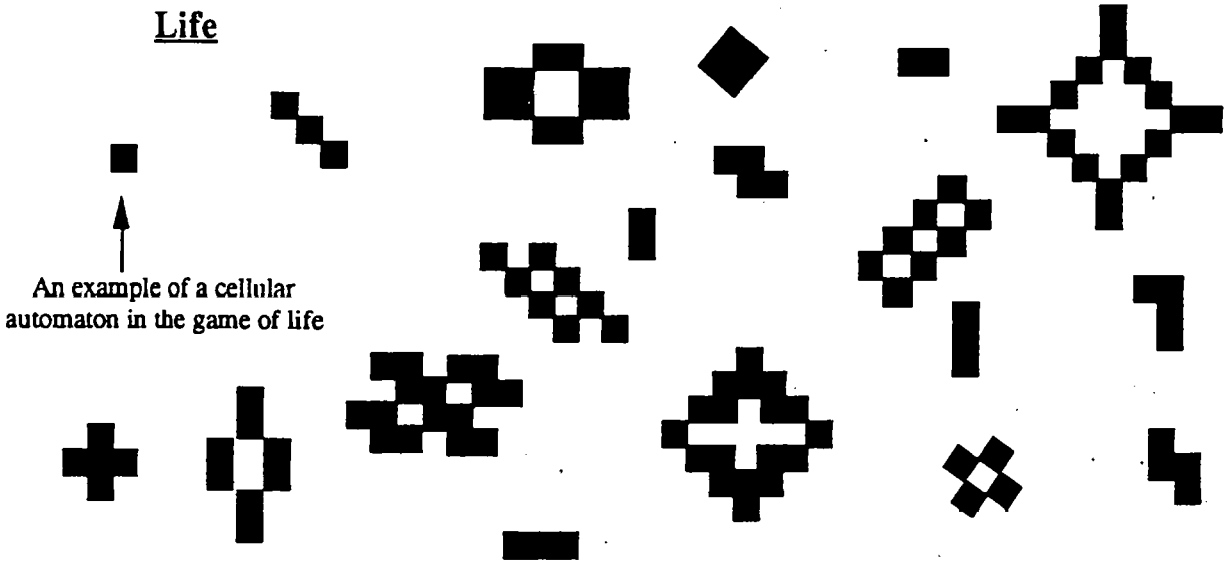

Figure 15

Fractals and cellular automata might be likened to ants and bees because although ants and bees have a fairly complex design and a small number of capabilities, their capabilities are increased many fold when they become members of a colony or hive. As the cellular automaton increases in number so does the complexity of its possible arrangements; similarly as the bee works as part of a hive it develops complex social behaviour and it is the interaction with other automata or bees, respectively, that dynamically increases the flexibility of the entity's potential actions.

There are, of course, differences between the behaviours exhibited by ants or bees and the actions exhibited by the cellular automaton or fractal. One of the most significant is that we would be loathe to describe the action of the fractal or automaton as 'behaviour' in the the ordinary sense of the word, since it is clearly not the result of the system being in any particular mental state at any time, nor because of the simplicity of the system can it be the result of a complex internal physical process. In the ant or bee, and even in inorganic systems of greater complexity than the fractal or cellular automaton, it is possible to say that the system behaves 'as-though' it understands, and for us attributing a mental state this is often all we demand as 'proof' of its existence. If this is then combined with our apprehension of the cellular automaton as a very simple system it would be difficult to ever say of it that it acts with any mentality or even any simulation of mentality. Thus, even though the fractal and cellular automaton break the 'rule' created by Figure 14 above, neither of them is likely to be attributed mental states on the basis of their action or of our apprehension of them. I will move on now to look at other areas of the diagram.

At the extreme top right-hand corner are human beings because they are, by general consensus, the most complex system of which we have a comprehensive, but still by far incomplete, knowledge. They have an extremely complex physiology, a complex social environment, complex relationships with other human beings and other entities whether organic, inorganic or altogether non-physical[8], they possess language and are capable of creation using symbols and non-symbolically. On top of this they are capable of analysis, again using symbols, but also incorporating non-linguistic gestures

such as facial expressions or remonstrating, for example, with a wagging finger or shaking head.

The position of human beings clustered alongside other higher-order primates is an interesting one for not only does it show that their architectural complexity closely resembles each other (as was previously stated in chapter four, section 4.2.1., ninety-eight percent of our DNA structure is identical), but also that they have a similar level of flexibility which means the other apes are capable of many of the things that human beings can do. They too have societies and complex social roles that each member must fulfil if they are to remain in the social group.[9] Not too unlike the "initiation" rites that potential gang members have to undergo, or the work quotas that have to accomplished if the employee wants to keep his or her job.

What then are the behaviours of which the human being is capable but the other apes not; in other words, why is the human being at the far right top corner of the diagram a little way beyond any of its closest relatives. There are the obvious physiological differences such as human beings stand upright, and from this accomplishment the earlier hominid earned the scientific name *Homo Erectus*, or "upright man". They had hands that could be used in defence where they had no sharp teeth or claws, the same hands could also make precise tools to hunt and carve up the spoils of the hunt. These sorts of differences might be classed as evolutionary since these are the changes that set human beings on course to become the species we know today.

There are other differences which are much less tangible and it is these that we more commonly associate with *Homo Sapiens,* or "wise man". They are things like having differing levels of consciousness, the possession of high-level mental states such as self-awareness, being able to communicate in a sophisticated manner so that one can speak of oneself in relation to one's world and *the* world whilst logically upholding a distinction between the two, and being able to consider abstract concepts that are from a superphysical world and not a phenomenal one. However, as we have seen in chapters three and four it is difficult to state where and when mental states such as

consciousness and self-consciousness start and stop. It would be asking for trouble to say, for instance, of a macaque monkey that it was unaware of itself for studies carried out around 1952/53 were able to show that macaques are capable of exhibiting extraordinary behaviours indicating quite a high level of awareness of both themselves and their surroundings.[10] One of the monkeys, Imo, seemed brighter than the rest and quickly saw ways around the difficulties that she encountered, for example, when given a sweet potato to eat Imo took it to a nearby pool to wash off the remaining sand and dirt; the other monkeys soon followed her example. On another occasion when rice had been sprinkled on the sand, and the other monkeys were carefully picking the grains of rice from the sand, Imo took handfuls of rice and sand to the water and threw them in, there the rice floated and the sand sank enabling Imo to scope the rice from the surface. Again, possibly seeing the usefulness of her action, the other monkeys soon followed suit.

Indeed it is hard to say just what Imo's behaviour shows. It cannot definitely be said that it is self-conscious or the result of a high level of awareness of both environment and herself in that environment. But what is possible is to say that there must be some element of sophisticated interaction between Imo and her world. Pin-pointing just what it is and what her actual mental states were is certainly a matter of continued debate. If the behaviour were exhibited by a human child it would certainly be considered to be the result of that child's prodigious intelligence. But there are three reasons why the child but not the macaque would be given the 'intelligent' benefit of the doubt; (i) we do not fully understand the mentality of the macaque, (ii) we do not credit it with the sort of intelligent behaviour we associate with ourselves or our children, perhaps because it would undermine our own superiority as intelligent beings,[11] and (iii) the macaque cannot explain to us why it does something for we have no shared language.

A second interesting issue that arises with this example is the matter of how the other members of Imo's social group were able to recognise her behaviour as a good or useful example to follow. It might simply be that they saw she was eating when they

were not, or that she might have been one of the dominant females and a troop 'leader' whose behaviour would be emulated. Whatever it can be put down to there is an element of understanding and communication that must also be present among the troop the makes it possible for the other monkeys to realise that they should take any notice of Imo's example. It is, however, doubtful that we could with our present technology measure the extent to which the capability to understand and communicate is present.

Going back for now to Figure 14 some more information that can be gleaned from it is that human beings have an extensive and potentially unlimited range of capabilities. There are some things that they are physiologically incapable of achieving such as self-powered flight, running as fast as a cheetah, or inhabiting the ocean bed, but these are things that human beings have managed to overcome using their intelligence, adaptability and physical capability for building instruments and machines to do these things for us. Human beings can now fly, run and swim to the bottom of the sea although their efforts are still not self-powered. No other system has yet shown that it is as capable of overcoming obstacles to its progress as the human being.

The other systems in the diagram are more obviously limited; the inorganic system by its design and the organic system by its physiology. Internal physiology and architecture have developed with the needs of the species as objective. An example of this might be the difference between bees and ruminants, such as cows or antelope. A bee does not have to digest its food twice in the way that a ruminant has to, but a ruminant does not have to function as a worker in a complex social hierarchy building cells, tending the larvae and feeding and cleaning the queen. They have different needs so they have a different structure with the flexibility to carry out vastly different tasks.

I shall now briefly look at how this diagram might be improved and then move on to examine another diagram which has been modified only slightly.

There are two main areas of concern in this diagram, the first is its lack of accuracy, and the second is that it has been possible to plot only a few systems. The latter of these two difficulties can be overcome by increasing the size of the diagram to allow for all the necessary information to be included, but the diagram would quickly become of

immense proportions. The other alternative would be to choose only those systems that are representative of a certain degree of complexity and a certain amount of flexibility and plot only those. This is what has been done in a very limited way in this diagram.

The former problem, that is, the failure in the accuracy of the diagram and the information it carries, could undoubtedly be improved by the axes being drawn more precisely so that degrees of complexity or particular capabilities might be stated explicitly. This, too, could be rectified by simply increasing the size of the diagram and adding in new gradations of the axes. However, the issue of accuracy is more likely to be improved by the addition of a third dimension which would allow points to be plotted more specifically, whilst also increasing the amount of available information. The result would be a three dimensional cluster diagram, that might be theoretically positioned in a fourth dimension of time as well.

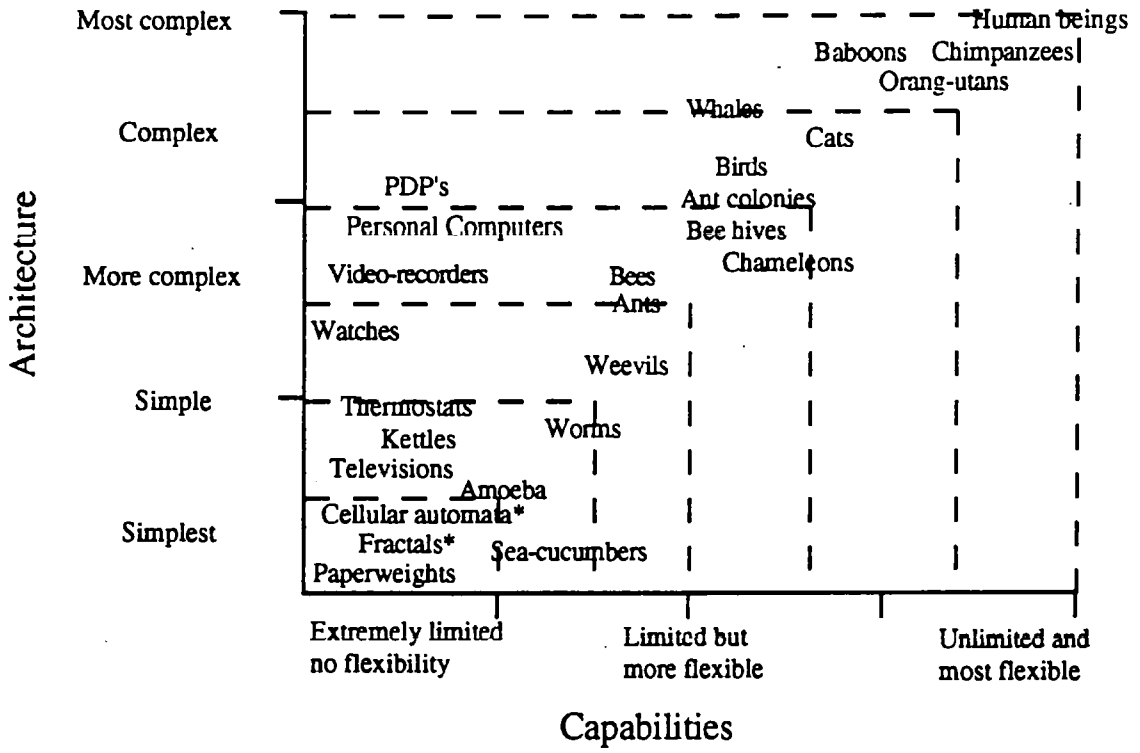

Figure 16

In Figure 16 only one slight modification has been made and that is the addition of dashed lines to more accurately indicate the points where architectural complexity and capabilities meet. So at the point where the highest level of architectural complexity and

the greatest flexibility of capability intersect we find 'human beings'. Again closest to human beings are others of the higher order primates. The rest of the diagram also remains the same with inorganic systems occupying the middle and lower left hand-side and organic systems taking up the central diagonal from the bottom left to the top right.

All the same successes and failures that hold for the first diagram hold for this one and there is one added difficulty which has been brought about by the introduction of the dashed lines. There is now the implicit suggestion that anything that is capable of the behaviour indicated by the intersecting of the two lines, upon which, or close to which, they have been plotted, is also capable of any of the behaviours that fall within the domain of those lines in the rest of the diagram.

So if we take cats as our example we can see that they are within the second set of dashed lines, indicating that their capabilities are still limited but that they have quite a high degree of flexibility in their repertoire of behaviours and a complex architecture. They are not, it seems, as complex as whales but they have a greater flexibility or range of possible behaviours then the whale. The position of cats on the diagram also tells us that they are more complex and more flexible than birds, colonies of ants, hives of bees, chameleons and many more. This may or may not be the case, but what is also suggested by their position is that they are capable of all the behaviours that the less complex and less flexible systems are capable. In reality this is not the case and again it is a problem brought about by the limitation of the axes, the choice of criteria upon which systems are to be plotted or measured, and the introduction of the lines that make the suggestion possible.

Another obvious example of this type of failure can be seen if we compare the capabilities of a thermostat with the capabilities of a human being. It is undisputedly the case that human beings are vastly more capable than a thermostat but when it comes to discerning subtle changes in the temperature of a room the thermostat, unless faulty, will win hands down against a system that has limitations in that particular respect.

It may simply be that it is too difficult to look at each system's architectural complexity as a whole. For a cat may be capable of better night vision than a canary but

the night vision of an owl must be at least on a par with that of the cat. So that the problem could be said to have been brought about by general nature of an 'architecture' criterion that makes no allowance for a specific aspect of architectural complexity that offers a particular species or member of a species, in this case the cat and the owl, a capability that another member of the species or a different species altogether does not possess.

Indeed to compare the complexity of a cat with the complexity of a horse or a human being might simply be perverse for they each have different functions, needs and levels of capability, flexibility and adaptability. One solution might be to plot classes of animals, such as mammals, birds, fish, invertebrates, and insects, and then in the same diagram also plot different types of inorganic system, for example, serial processors, neural networks, simple binary switches, video recorders, and so on. A comparison of this sort might be feasible, and perhaps even favourable if there were also the addition of a third dimension against which the system's adaptability to survive within a changing environment could be measured.

I shall look at one more two dimensional representation before moving on to construct and examine a three dimensional model of the relationship between architecture, capability and adaptability.

There is a change of form in the third diagram. No longer are there any axes against which the systems can be plotted. Instead there are a number of rings that vary in size as an indication of the complexity of the system. Thus the larger rings belong to more complex systems than the smaller rings because the system's represented by the larger rings have the potential to possess more mental or internal states.[12] The outermost ring represents all the mental or internal states possible for all types of systems. This might also be described as the largest 'state space' for it is the space in which all possible states are contained. The overlapping of the rings indicates where there might be some coincidence of mental states, for example, that all systems can be aware to some extent of their environment, or that every system is capable of at least some degree of intentionality, (at least according to Dretske).

It is a diagram that is difficult to understand because a lot of its information is carried implicitly and not stated explicitly as it was before on the axes of the other two diagrams. We have merely a vague idea about the complexity of the system which is founded only on the size of the ring in which the system is contained; and no mention is made of capability except that we can extrapolate from a system's having a wide range of possible mental states that it is also capable of exhibiting many different types of behaviour.

One of the overall problems of Figure 17 is that it seems to be attempting to present too much information at one go and the result is that all the information it carries, which is in fact an immense amount, becomes blurred. This diagram carries no more information than either of the other two but their method of representation was plainer than this which is simply designed to show overlapping relationships between categories of things.

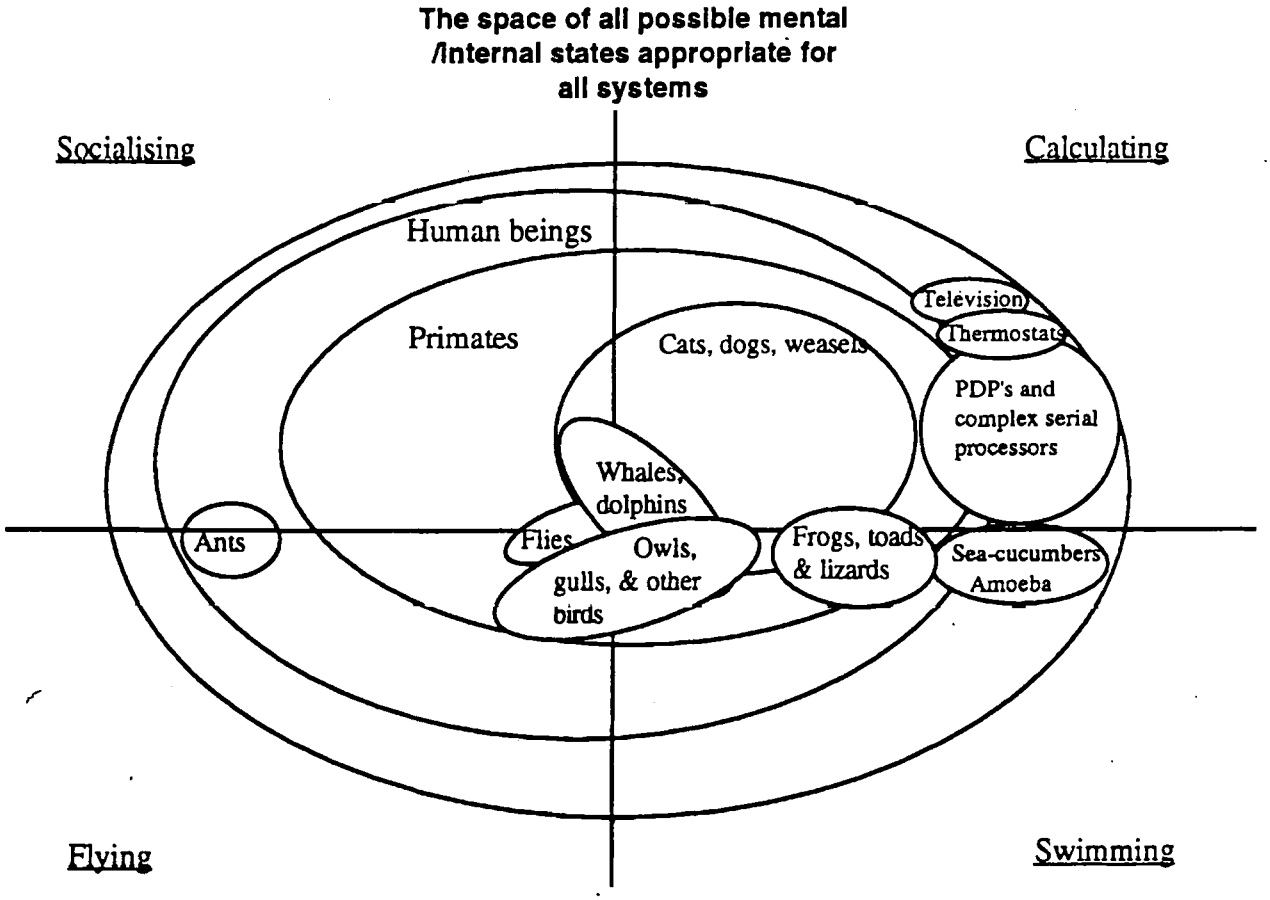

Figure 17

In the other two diagrams systems are plotted as points on a graph with no mention of overlapping states, except to say that those systems that are clustered together are more likely to have similar states, or commonalities between their internal states, than those that are dispersed in other areas of the diagram. In fact it might be argued that less information is carried in this diagram than in either of the other two, but I would counter this by saying that the difference is not between the level of information content, but rather between two ways of carrying information, explicitly and implicitly. In the figures 14 and 16 the information is stated explicitly, whereas in the this diagram the information is present in a more implicit form.

But there is another problem; and it is this: because we in fact know very little about mental states it becomes almost impossible to say where any commonalities of mental states really do occur. The suggestion then is that the information that the diagram carries in explicit form, that is, the very obvious overlaps between different systems, is so vague that it ceases to be informative. After all the most important aspect of a diagram should be that the information it carries is explicit or up-front and this diagram does not contain very much explicit information, and that which it does contain is dangerously over-generalised, running the risk of telling us nothing of either use or interest.

Again, because of vagueness of the diagram I am compelled to conclude that any attempt at plotting individual instances of a species would not be a wise thing to do with the limitations of space in the diagrams and the need, at this stage to establish some overall picture of the relation between different systems, their possible mental states and their behavioural capabilities. In the next diagrams it will be more sensible, from a perspective of increasing the available information and accuracy, to compare classes and types of systems. Too much information has to be deliberately left out if only one or two examples from a particular class or type are plotted. So that, subtle differences, for example, in architecture cannot be reported because of the lack of space and the overwhelming number of species and members of those species which we might choose to examine.

In the next section I will look at the advantages of a three dimensional representation. It will be a cluster diagram in the style of diagrams one and two but with the addition of an axis to ascertain the 'adaptability' of the system to respond to change. So now, classes of organic systems and types of inorganic systems will be plotted alongside their capabilities, architecture and adaptability. This new procedure has the potential to increase the amount and accuracy of the information that will be offered.

6.3.2. A three dimensional model

Although in this section it is my intention to achieve a greater degree of accuracy, it should be said that the representation is still by no means complete, and the reason that this representation, and indeed no representation of this sort can ever be complete is because no perfect set of axes exists against which information can be measured. No perfect set exists because the information we are looking for will always depend upon what it is to be measured against and we are limited with the spatial structure of the diagram to measure at most three spatial dimensions, and if it is then plotted through time, one temporal dimension as well. For this reason some information will always have to be omitted or generalised to fit the axes that we wish to be present.

Now to the question of what such a representation would look like. I have said that the ascription of mental states is a product of the apprehended complexity of a system and its ability to behave in what we consider to be a human-like way, now what we want to look at in this same context is the relation between complexity, capability and adaptability. Capabilities are what are implicitly ascribed when mental states are attributed. For instance, if I say of Arthur that "He believes that the idea of curved space is open to fundamental misconceptions" I am ascribing to him a great many complex capabilities. Not least of these is that he can understand what I am saying and form beliefs about it. For each different system the environment and that system's behaviour will change thus the capabilities that are attributed to it will also change. Thus the best way to start to build up a three dimensional representation might be to begin by

altering the the two dimensional axes so that 'complexity' and 'adaptability' are measured first and 'capability' is then superimposed as a function of them.

The diagram will take on the following appearance because I am arguing that capabilities are a result of the system's complexity and its ability to adapt to new and continually changing stimuli within its environment:

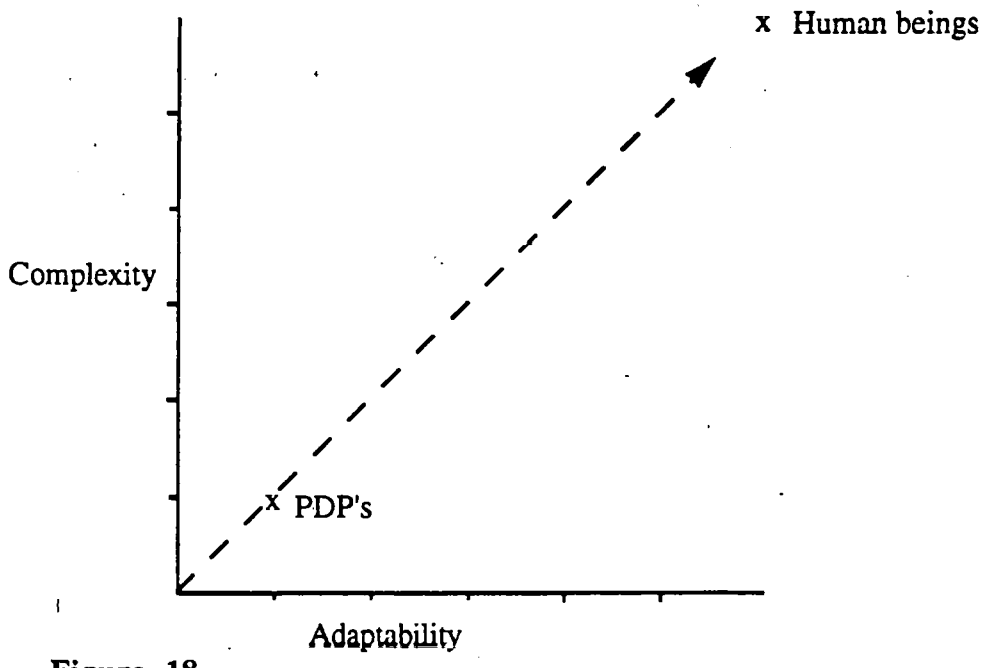

Figure 18

It can now be argued that capabilities are dynamic whereas adaptability and complexity are static things that are either present within a system or not. Of course, a proviso must be added, and it is this; 'complexity' and 'adaptability' are two characteristics that are present in some machines. The former is necessarily related to our understanding of different kinds of machines, for I am sure that not so long ago a thermostat would have been counted as a complex machine, but with our technological advances it has been relegated to the realms of simplicity itself. The latter, that is 'adaptability', is something that has been incorporated into Parallel Distributed Processors or "Neural Nets" so that they can exhibit learning behaviour and thus adapt to changes in their environment. This is not to deny that the thermostat is adaptable, for it is, but its ability to adapt is extremely limited for it has only a three possibilities, 'on', 'off' or 'no change'. Because of its simplicity and the strict limitations set on its ability

The next step is to begin to superimpose the third dimension of capability as a function of the two already present axes. One way to do this would be to outline each of the separate groups of crosses and draw columns from the edges of the outline to the *x* or bottom axis thus producing a type of three dimensional bar chart. A second method would be to select a couple of crosses from each grouping and make them three dimensional by again drawing them as columns to the bottom axis. This would make the 'bars' thinner and more easily differentiated from each other. A third possibility would be to divide the diagram up into, for example, 144 squares, and to then plot the positions of every cross within those squares. This could then be plotted into a three by three matrix and when processed mathematically a three dimensional diagram formed.

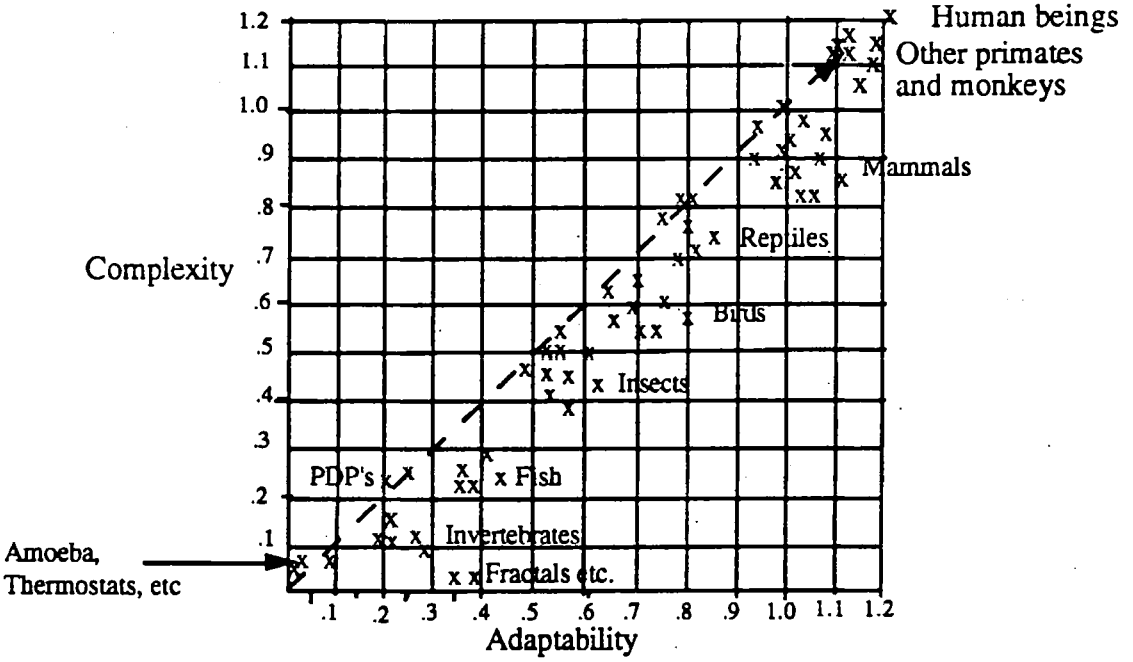

Figure 20

Circling the crosses and creating columns quickly becomes unfeasible on a diagram of this size for the lower levels then become unreadable as columns. For example, when trying to create a column from the class of invertebrates there is no room available in which to make the encircled class three dimensional. In other words, the species that lie along the bottom axis, the simplest systems, cannot be raised to a sufficient level of complexity to show up as points on a three dimensional diagram. Similarly for drawing

to adapt the thermostat would be positioned in the diagram above at the very bottom left hand corner where the two axes interconnect.

Entities such as fractals and cellular automata would have a more obvious presence on this diagram for they, being simple yet very adaptable, would be half way along the adaptability axis and in the lowest position on the complexity scale. Again they appear as exceptions for they sit outside the general trend of the graph which follows the central diagonal line leading from the lower left corner to the upper right corner.

If we move on at this stage to plot a number of different species alongside some inorganic systems on the new set of axes we should get a diagram that would conform approximately to the following, where the general trend is becoming more apparent.

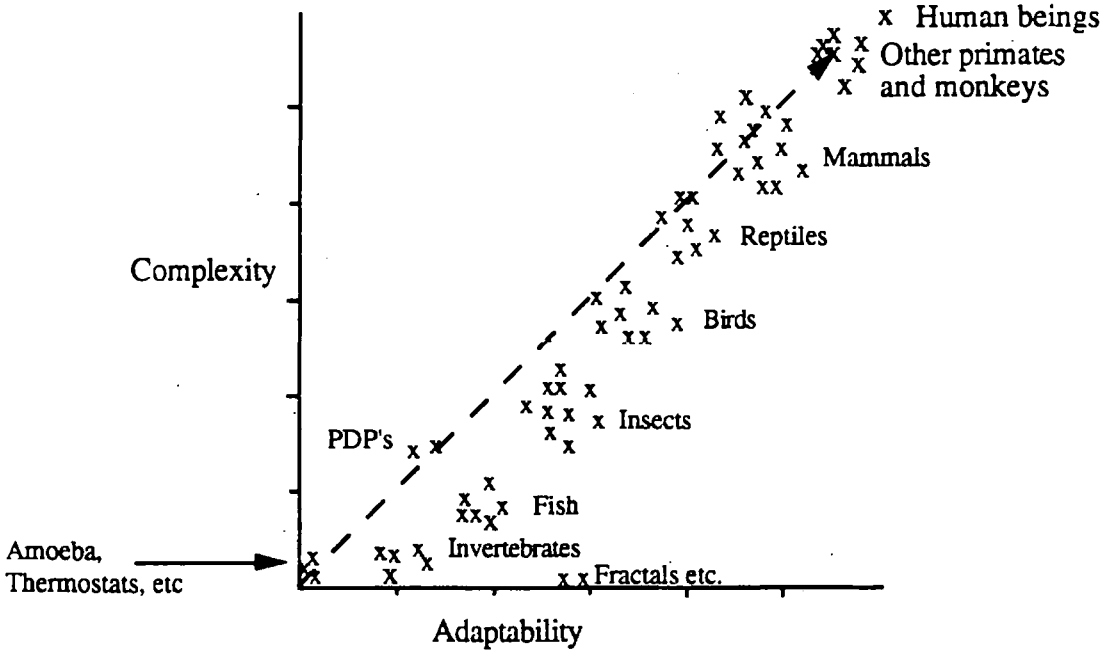

Figure 19

The different species and systems will remain in approximately the same positions as they did in all of the previous graphs even though one of the axes has been changed. That there is very little difference in the respective positions of the clusters is mainly due to the fact that the term "flexibility" also contained the implicit meaning of "flexibility to respond to changes in the environment" and thus being capable of a great many things, now the term "adaptability" explicitly covers both notions whilst leaving "capability" to be examined as a separate issue.

columns in relief from a selected set of points to the bottom axis would be workable on a much larger diagram but on one of this size too much information is unavoidably omitted. Drawing a grid or framework into which the points are plotted, or indeed which could be placed over the existing points is some sort of "mid-way" method between these two possibilities and Figure 20 shows the beginning of just such a procedure.

Although this diagram is still only two dimensional the addition of the grid or frame permits us to see that each of the crosses or points occupies a specific value that can be measured to within a tenth or a hundredth of a decimal place. For instance, the values for each of a selection of points in their respective classes would be as follows:

```
Susan'sPoints = Table[{LabelText["Human Being", {1.0, 1.2}],
                       Point[{1.2, 1.2}],
                       LabelText["Primate", {1.05, 1.06}],
                       Point[{1.18, 1.15}],
                       Point[{1.12, 1.17}],
                       Point[{1.12, 1.12}],
                       Point[{1.09, 1.12}],
                       Point[{1.15, 1.05}],
                       LabelText["Mammal", {0.97, 0.97}],
                       Point[{1.11, 0.86}],
                       Point[{0.99, 1.0}],
                       Point[{1.08, 0.95}],
                       Point[{1.03, 0.98}],
                       Point[{1.07, 0.9}],
                       Point[{1.0, 0.94}],
                       Point[{1.02, 0.88}],
                       Point[{0.99, 0.91}],
                       Point[{0.93, 0.90}],
                       Point[{0.94, 0.97}],
```

```
LabelText["Reptile", {0.8, 0.78}],
Point[{0.86, 0.74}],
Point[{0.81, 0.81}], etc.
```

Using these co-ordinates it is possible to plot this graph[13], although still only in two dimensions, to a much higher degree of accuracy using a mathematical tool such as *Mathematica*.[14] The clusters of points will be used by *Mathematica* to form the more accurate graph as seen in the diagram below:

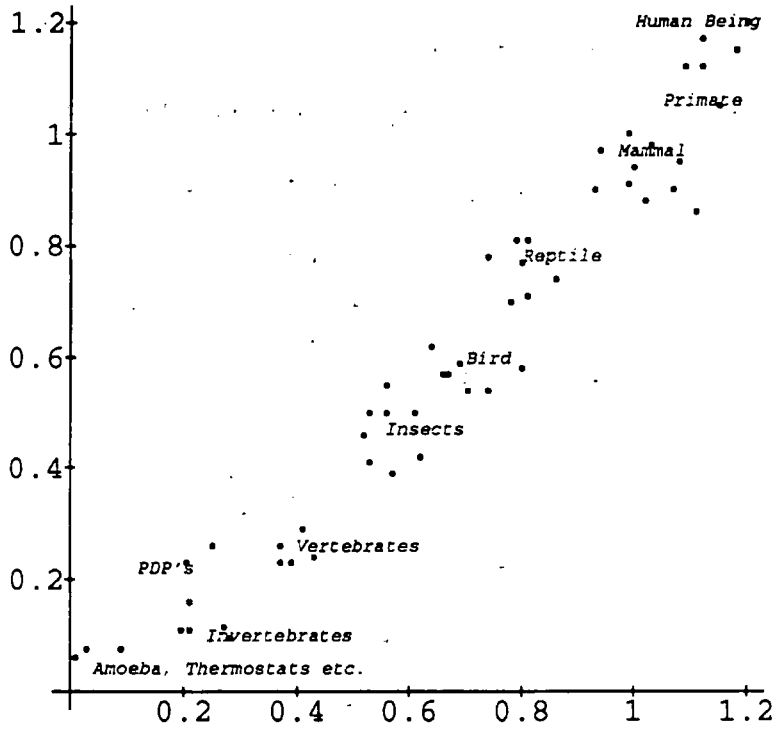

Figure 21

The next step in the procedure is to add a third dimension that can be estimated and plotted in the diagram, thus bringing the points into relief. This new dimension will be the capability of a system as a function of the complexity and adaptability of that same system. To enable us to remain within the extent of the other two axes, still using the same system and range of measurements, I will take the mean or average of the other two axes as representative of the value of a system's capability.[15] This being the case we can see using only a couple of possible data points that human beings, for example, would have a capability value of 1.2 since the values of both their x and y axes are 1.2 also; and a reptile with $x = 0.86$ and $y = 0.74$ would have a capability value of $z =$

0.80. Thus with three sets of axes our diagram shows a more definite and by now wholly determined relationship between the complexity, adaptability and capability of each of the systems. Our representation now looks like what Kevin Kelly has described as a *Possibility-Space Notation.* His diagrams are very similar to the following one and he uses them for essentially the same reasons; it is "a visual notation to render a simplified conceptual view of complex things".[16]

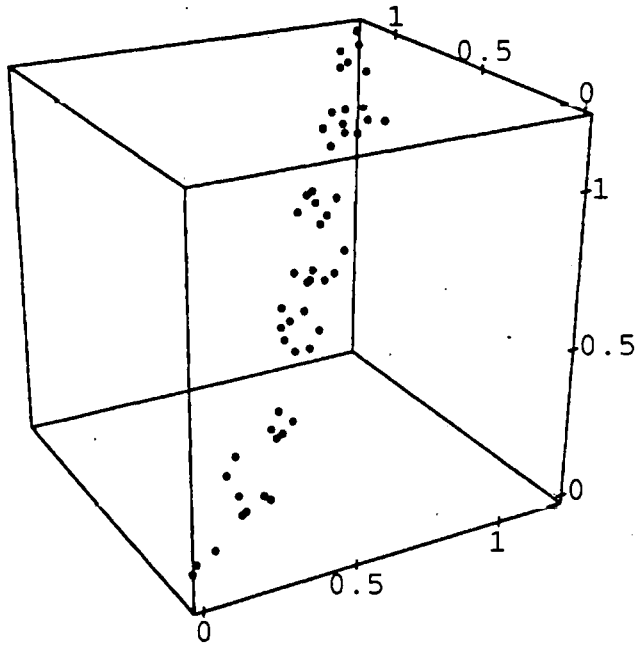

Figure 22

The third dimension has been introduced but the diagram is still very unclear. What we need to do is give *Mathematica* a function with which it can relate the points of data to every other point and also in relation to the rest of the possible points, though not explicitly estimated, covered by each of the three axes. This has been carried out in the next diagram and the difference is remarkable. The diagram now shows a more clearly focussed and purposive distinction between different types of systems or species. Indeed the diagram now begins to resemble a mountainous region with, it would seem, human beings having climbed to the apex of this, what could perhaps be described as, "evolutionary mountain". So human beings occupy the peak position because of all the

systems we know or create the human system is the most complex, with a complex physiology and a complex social environment. As a species human beings have adapted extremely well within a dynamic environment for they have made full use of their skills at communication and co-operation to share and successfully complete tasks. Instead of being dictated to by their environment human beings have learnt to a large extent how to control their environment to best suit them.

With greater resolution the dimensions and scale of the diagram are more evident and what we have really does look like a range of mountains with one prominent peak and a number of smaller hills with their own lower peaks. The most dominant of the peaks is where the greatest complexity, the healthiest adaptability and the largest number of capabilities come together as being characteristics of the human system. As each of these three characteristics diminish we move down the range to reach other simpler and less capable systems indicated by the lower peaks.

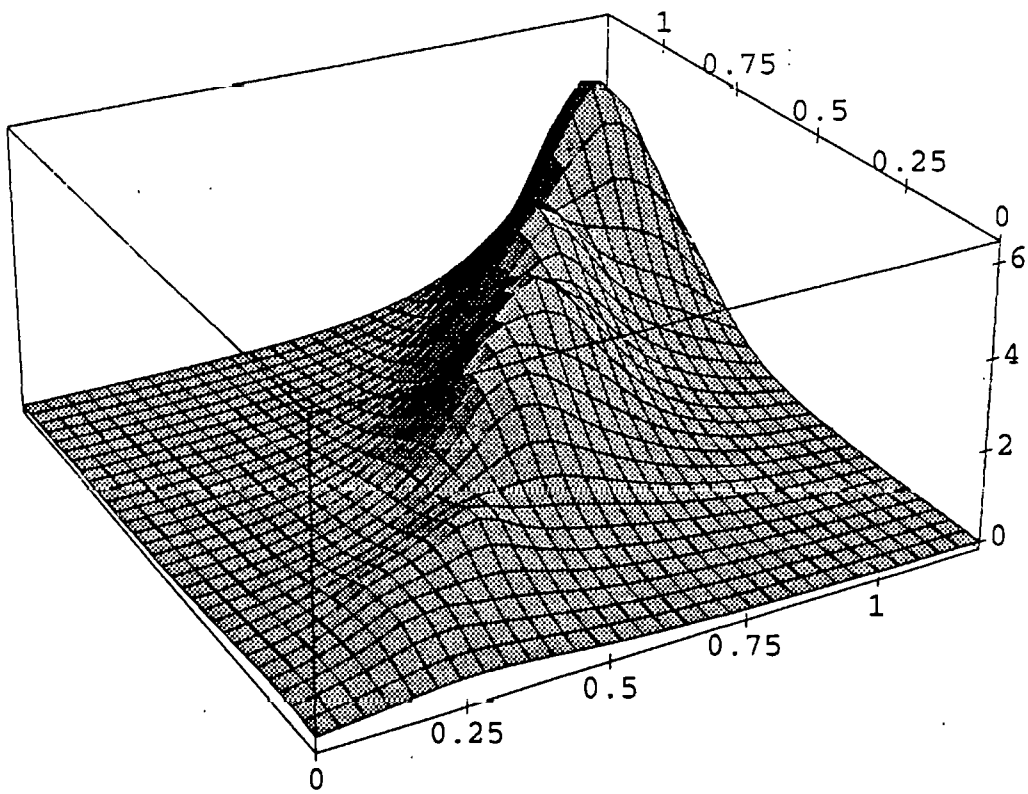

Figure 23

Next on the range, in fact still on the same 'mountain' of the range and just below the highest point occupied by human beings, are the other members of the higher order primates. There are a great many of these, ranging from apes such as chimpanzees and gorillas, monkeys such as macaques and gibbons, and lemurs such as the *sifaka* and *indri*. Of these three the apes are most often described as 'anthropoid' because of their many similarities to human beings. They are the closest living relative of the family '*Homo Sapiens*' and because of this they are capable of many of the things of which human beings are capable. For example, their faculties of perception are largely the same as those of the human system, as are their facial expressions with which they register intentions of, for example, friendliness or aggression. Indeed, as we draw our lips back over our teeth in a smile to let another person know that we mean no harm, so too a young male chimpanzee will employ a similar facial expression to convey to the dominant male of the troop that he does not intend to threaten his position in the hierarchy. Of course, there are notable exceptions to these rules when it comes to human beings, and one need only look to Shakespeare's *Hamlet* for the evidence:

> Hamlet: O villain, villain smiling damnèd villain!
>
> My tables - meet it is I set it down
>
> That one may smile, and smile, and be a villain.[17]

In this short section one great difference between humans and other apes can be seen to exist, and it is this, the human ability to deceive. For human beings can behave in one way that gives the impression of meaning a particular thing when in fact it is their intention to double-cross the other person who accepts their behaviour for what it is. This deceptive behaviour is not something that the other primates have acquired for their societies run on a much less complex basis where actions have a "face" value and are taken to mean what they state.

In terms of complexity and adaptability there are other differences that occur between human beings and the rest of the primates and these for the most part tend to be things such as the extent to which our society and our interactions with one another have evolved. So that our society, as human beings, has become a great deal more

complex than those of the other primates. One of the biggest distinctions can be seen in our ability to form language and use it as a shared tool, for it is this that has made it possible for us to conceive of ideas, discuss them with other people and transform them so that they can be put into practice. The complexity of our ideas reflects the complexity of our society and the sorts of knowledge we possess. Our knowledge can be expressed by us as individuals but is more likely to be expressed as a very small part of a larger whole which is our society and culture. Putting our ideas into practice is a reflection of the immense creativity of mankind and our creativity is just one very tiny aspect of our wide range of capabilities. It is true that chimpanzees are also being creative when they make tools but their tools are, as yet, very limited and it seems that there is no obvious element of abstract thought present in the process. However, this is not to say that their needs, and consequently their tools, have reached the end of their evolution, for we cannot ever know that. What it is possible for us to know is that it is at very best unlikely that their, and our own, evolution will now be at its end.

The rest of the diagram shows that a differentiation exists between all of the other species on the grounds of complexity and adaptability as well. The general trend is that the simpler the system the less adaptable it will be and the fewer capabilities it will possess. This is certainly the case even if considering only one variable, that of changes in temperature, and looking at the differences in the capabilities of mammals and reptiles. Mammals can adapt to even quite extreme changes in temperature, continuing to hunt, forage, and even play. They have the capacity to wake up quickly becoming alert and active in a matter of moments to any possible predator or prey. Reptiles, on the other hand, do not adapt easily to changes in temperature. They favour warmer climates than most mammals because they need the sun to keep them warm and they have no fur that they can fluff up to retain body heat. When the temperature drops reptiles, such as marine iguanas become lethargic so that the food they have eaten can be used to maintain a general bodily homeostasis. Similarly in the morning when the temperature is low the iguanas wake up only slowly trying to use up as little energy as possible and being slowly rejuvenated by the warmth from the sun. Even when the

weather is hot enough much of their day is spent either feeding or basking in the sun. They do not need to play and learn by that method to defend themselves against attackers for they have very few natural predators.

The flow of the diagram steadily continues approaching the bottom left-hand corner where the systems have the simplest architectures, occupy the simplest environments and exhibit the simplest behaviours. Such systems do not need to be very adaptable to continue to exist in their environments for very little is required of them. The thermostat, kettle, television, sea-cucumber and amoeba need only to be adaptable within their range of necessary functions.

One of the more interesting sections of the diagram can be seen around the intersection of $x = 0.25$, $y = 0.25$ and $z = 0.25$. This is where Parallel Distributed Processors cause a blip in the trend or flow of the organic systems in the diagram. PDP's are complex, but not strikingly so, for they have a more complex architecture than any other inorganic systems, but are only barely as complex as any of the members of the invertebrate family. Their environment is quite complex and they are capable of learning which enables them to adapt to new information that will be the cause of subsequent changes in their actions. Being capable of taking in information, and selecting those pieces that are most relevant to it, is a significant feat for an inorganic system to be able to accomplish and it is this 'accomplished' capability that has caused the PDP to interrupt the trend of the diagram, thus bringing an inorganic system into the relief of having a significant third dimension.

However, their range of possible actions is still quite limited for there is a lot of information in the world to which the PDP cannot respond. But this can also be said to be true for a great many more complex entities since all systems will be restricted by the natural limitations of their perceptual apparatus. However, on the grounds of complexity a distinction can still be maintained between PDP's and other systems for the more complex a system is the fewer will be the restrictions on its perceptual capabilities. So that a reptile, such as our marine iguana, possesses a much more

capable perceptual apparatus and there is much more in its environment to which it can respond.

At this stage it should be remembered that as we are talking broadly in terms of the complexity of architecture and environment, the adaptability to incoming stimuli, and the capabilities that are afforded to a system, the information we have will remain quite general. But this is no bad thing for it has got our "picture" of other systems into a more realistic perspective enabling us to answer the question "Where do we go from here?".

The next stage will be to consider some of the work carried out by Aaron Sloman that deals with the "design space" of the system. This work is of particular interest because Sloman emphasizes the need to look at the whole system if we are to have any hope of ever having a full understanding of the inner working, both physical and mental, of that system. He maintains that a great many of the drawbacks or obstacles that have been encountered by work carried out by Artificial Intelligence are due to an overwhelming concentration on the very small aspects of the system whilst continuing to ignore the system's architecture as an interacting, interdependent whole.

6.4. Design space

In his inaugural lecture to the University of Birmingham[18] Aaron Sloman argues that to make any progress in understanding the mind we first need to know two things; (i) what an intelligent system would need to be able to do for it to function and be described as intelligent, and (ii) what various mechanisms can already do that make us believe and attribute intelligence to them. Once we possess an adequate understanding of these two things we ought to be able to discuss mental concepts, such as 'intelligence' and 'consciousness' more successfully than we do at present.

The particular approach that Sloman proposes is a "design-based" one that looks at the mechanisms and architectures of a system as part of the space of all possible designs. Such an approach "requires understanding which features are important for which capabilities, and how the capabilities would change if the design were

changed".[19] The idea being that if we understand what capabilities a system possesses by its being designed in a particular way then it might also be possible to know how those capabilities would alter if the system's architecture or internal mechanism were to be changed in some way.

To begin with we know that between species there is a great deal of functional variation that is brought about by the vast differences that exist in the architectural, environmental and behavioural complexities of each distinct system. We also know that among the same species a great deal of functional variation exists because of the great complexity or 'richness' of their architectures. This is especially, and certainly more obviously, the case for human beings.

Many of the functional variations that exist between and among species or systems are brought about by the fact that the system possesses a potentially changing architecture. That is, an architecture that is dynamic in the sense of being able to adapt to incoming information that it has not experienced before. Thus in the diagrams above anything that is both complex and adaptable and showed up in the third dimension will have a dynamic architecture and be capable of adapting to suit changes that affect it within its environment. Sloman argues that one of the capabilities that changes through learning is perception; for example, our ability to recognise and interpret three dimensional shapes and different sorts of motion. I would go a stage further and add to this that, as higher order primates, all the capabilities of the human being have, and still continue to, change through learning which allows for adaptations in both our architectures and internal mechanisms. For example, as has become necessary we have become bipedal and an arch has developed in our foot to make running easier. We have developed binocular vision because with the absence of teeth and claws as weapons our only advantage was to see the predator or prey advancing from a long way off. So the internal architecture of the human system has adapted quite strikingly to changes within its environment. And other systems have also evolved and adapted for each of them has had to fit into an ecological niche in which it has been possible for them to survive.

On the more immediate scale changes in the environment will not bring about alterations in the architecture of the system; however, on this scale changes can be seen within the system's internal mechanism. One good example of this might be the secretion of adrenalin by the endocrinal organs when the organism is frightened or excited thus causing constriction of the arterioles, dilation of the pupils and acceleration of the heart. It is also known to concentrate the mind on the object that is the cause of the excitation. A second example, is that of the emergence of capabilities that are unexpected just as connectionists argue that intelligence or even consciousness will be an emergent property that arises out of the creation of a complex enough parallel machine.

So neither the architecture nor the mechanism of any complex system, that also has the ability to adapt to changes within its environment, can be said to be static. Its internal mechanism must be capable of changing to aid its survival and it must be capable of assimilating these changes into its overall architecture if they are going to continue to be necessary in the future. In this way the internal states of a system are always going to be in some degree of fluctuation, although for the systems chemical changes these may for the most part be measurable. However, when it comes to emotional states, that may to some extent coincide with the lack or secretion of certain chemicals in the body, we find that they are much more difficult to measure for they are similar in form to mental states, such as 'knowing', 'believing', and 'wishing'. Their essential vagueness and intangibility does not deter Sloman for he talks of "Hierarchies of Dispositions"[20] and I foresee that he will have problems similar to those experienced by Dretske and others who have tried to stratify mental states. I shall look more carefully at how Sloman attempts to make his dispositional divisions.

6.4.1. "Hierarchies of Dispositions"?

In this instance Sloman is primarily concerned with the human system and he claims that two sorts of dispositions exist, those that are long term and hard to change and those that have only a short term existence and are episodic or transient in nature.

Of the first type he cites personality traits and attitudes, which I suppose to be dispositions of the following kinds of nature "happy-go-lucky", "conscientious", "sad", or whatever, all of which make up the underlying nature of the system.[21] The episodic dispositions are things such as moods, beliefs, desires, and intentions which can change from one moment to the next depending on the social context of the individual.

Superficially the hierarchical division seems to be fine except that surely it is not realistically possible to make the two types of disposition appear so distinct for the long term dispositions are necessarily made up of a mass of interweaved and interacting short term dispositions. In turn the short term dispositions must also be strongly influenced by the firmer more lasting personality traits that may even seem to have been inherited from a parent or other relative, as in "You're just like your father!". If there is any overlapping between the dispositions their distinction will start to blur and they become the sorts of states that we have already found it hard to discuss because we cannot define or describe them and the hierarchical tools that have so far been used in an effort to define them have been seen to be of little use.

Later Sloman decisively announces the need for a new vocabulary or conceptual framework for discussing mental states and processes and the commonly ill-conceived notions of 'consciousness' and so on. Only here do I find myself beginning to agree with him. He argues that our vocabulary for such states will improve alongside our understanding of the relevant mechanism, and I agree that this cannot but be so for the more information we have the greater will be our understanding and with an increased understanding we will develop an enhanced vocabulary with which to handle our new knowledge. However, I do not agree with him that it is an understanding of the system's mechanism that will yield this fuller and more useful vocabulary. There are other things that are also necessary; firstly, to look at the system's overall architecture, secondly, to examine the rôle that the system plays in a wider and more universal type of architecture, and thirdly, to look at the reasons that a system has a particular form or

architecture and then look at what it is capable of doing because of that architecture that other systems are unable to do.

There are, it seems then, a great many things that are omitted in Sloman's consideration of how the architecture and mechanism of a system relate to the proper functioning of a system, but perhaps one of the more interesting aspects, that I have so far rather purposely overlooked, is his emphasis on a "design space" in which different systems with different architectures occupy different points. I shall now look at in more detail.

6.4.2. Dispersal across the design space

Sloman argues that if a system is to have a number of different capabilities then it needs to occupy a multiple of points in the "design space". The implication of this is that if a system can only carry out simple activities it will only need a simple design space and correspondingly if it is to be capable of more complex things its design space will need to be more complex.

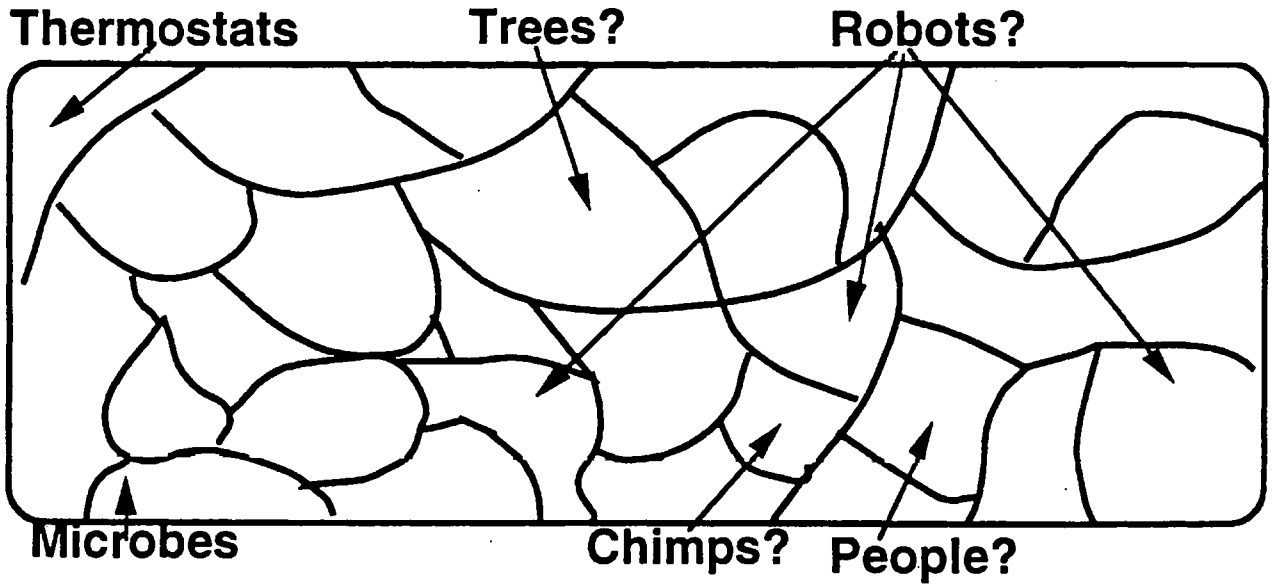

Figure 24

Human beings are capable of carrying out a rich and complex array of tasks and Sloman infers from this that they also occupy a rich and complex design space. So far so good, for I too would agree that the macaque is capable of a good many things and it too will have a rich and complex design space similar in many ways to that of the

human being. Likewise almost at the other end of the scale an invertebrate, such as the sea-cucumber is capable of very little; in my diagrams it occupies the lower left-hand corner and in Sloman's design space it has only a very simple area.

Another interesting aspect of this work is that Sloman concludes that the richness of a system's complexity is not enough for it to be capable of exhibiting complex behaviour, Sloman also demands that we look at what the system needs to sustain its existence in the environment it occupies. This is most likely to be its ability to adapt and respond to change in the most favourable way for its own survival, thus making it a dynamic system.[22]

While similar in some respects, the general approach of Sloman's work seems to be quite different to mine for he is interested in the engineering aspects of designing a functioning mind. His proposal is for an explicit investigation into architectures and mechanisms on an individual basis, whilst I propose that we ought to first establish a structure from which we can start our investigation on a more global scale looking at systems and their interactions rather than individuals. Nevertheless, by travelling along different routes we have drawn many similar conclusions, for example the importance of the sustenance of the system, concentrating on the system's ability to adapt to changes when necessary in order for it to survive. These I will now discuss in section 6.5, the final part of this chapter.

6.5. Conclusion

I will now draw together the main conclusions of chapter six and briefly discuss the capabilities of a number of different systems whilst also looking at what mental states are required by those systems for them to be capable of their own particular sets of actions.

The three things that have emerged as being of great importance. The first emphasised the uniqueness of every chosen axis or set of axes for whatever interpretation we make of something it will always depend upon what we chose to examine or measure it against. Thus the information that we have obtained about the

presence of mental states has depended upon our looking at them in connection with the complexity of the system and its environment, its adaptability within that environment and the capabilities that arise from the degree to which the system is complex and adaptable. Had I chosen to examine these systems in relation to their use of language, their ability to use tools, or their ability to form and act as part of a social group, the picture would have looked very different. Indeed many systems would not have made much of an appearance, for example, neither thermostats nor PDP's form social groups and because of the immensity of their size and a need for large quantities of food each day just to sustain them, orang utans have to live a solitary life.[23] So orangs too would not have appeared on a diagram that set out to investigate systems in relation to their ability to form social groups. Thus the criteria by which we choose to define a thing(s) are very important since even members of the same family can be excluded from their own family description if a new set of axes or even a slightly different definition is employed. A good example is the whale for it can be defined as a fish if the criteria are that it dwells in the sea and swims with the aid of fins and flippers, however, if the criteria are that it is warm blooded and produces milk with which to feed its young, then it is defined successfully as a mammal and not a fish after all.

The second conclusion to be drawn from all of this is that two dimensional representations are limited by their size and subsequently by the quantity and accuracy of the information that they can provide. The simplest way around this is to increase the number of dimensions thus enabling information to be plotted, not only in relation to two sets of interrelated axes, but now in a much more complex tripartite relation. Just as Schubert's "Unfinished" Symphony while sounding good when played on one piano, sounds superb when played by a whole orchestra in a good auditorium. The music becomes enriched with an increase in participation by other members of the orchestra so that different sections will be brought alive by the strings, the woodwinds and so on. In much the same way the information has become enriched in the new diagrams by the addition of another dimension that can offer more information and a fuller perspective of what is already there implicitly. For the amount of information I

am trying to portray, that is, that a relationship exists between mental states, architecture and capabilities, increasing the dimensions is absolutely necessary.

The third conclusion that can be drawn from this chapter is that using a taxonomic method we have been able to show a comparison between the differing capabilities of a wide range of systems of both the organic and inorganic varieties. It has meant that at last it is possible to compare and contrast the differences between the complexity, adaptability and capability of a number of systems and that these systems can have different forms and different functions, so that the capabilities of a cat or a lizard can be compared with the capabilities of a thermostat or a more sophisticated machine such as a PDP.

Complexity was examined earlier and found to exist in many subtly interacting ways, architectural, behavioural and environmental. These have been brought together into a single entire concept of complexity that has made our task of drawing systems together on a basis of their overlapping similarities a lot easier. Working through these diagrams and amending them as I have gone along it has become possible, using the same criteria for each, to look at and compare machine states and mental states. This had not been possible before because mental states are vague and not easily differentiated whereas machine states are easily defined and quantifiable. Thus it was possible to differentiate between and stratify the states and functions of a machine, but not to do a similar thing with the mental states of a living system. Using this method we now know that for looking at vague concepts such as mental states cluster diagrams are a great deal more useful than the less flexible form of representation offered by a hierarchical stratification.

At last we are in a position where we can say with a greater degree of confidence which mental states are necessary for system to occupy a particular position in relation to other systems on the diagram. I shall bring this chapter to a close by examining this issue.

6.5.1. Which mental states are necessary for which capabilities?

Human beings are the most capable systems that we know. They occupy the top level of the all of the cluster diagrams, whether the diagrams are two or three dimensional. The question to raise now is "What has elevated human beings above all other systems to this level?". Well, it is certainly a matter of "complexity" as we have seen in chapters four and five, and such complexity is not just a matter of our being internally complex organisms with complex architectures and mechanisms, but is also an issue when we consider the complexity of our societies, our often very different cultures and the sorts of complex behaviours we exhibit daily just in the course of living.

On top of this we are extremely adaptable when it comes to analysing new information and adjusting our own behaviour, and even sometimes the behaviour of others, in accordance with new information. Our perceptual field is vast and we take in a lot of information that is superfluous which we then ignore or store for later use. We try to understand the information that we have selected for attention and from this new knowledge we can form new beliefs, feel satisfied that we have more evidence to maintain our old beliefs or bring our old beliefs into alignment with the new information.

So because of our immense complexity and our ability to adapt and survive we find ourselves to be capable of a great deal of interesting and useful behaviour. We have splendid perceptual skills with which to take in information about our world. We can understand, make judgements, store our new information as increased knowledge, and form beliefs about it with which we can deal with the world in a more sophisticated manner than we could before.

Our existence is largely egocentric, as, it must be said, is the existence of all other animals. One question that still looms large is whether other animals know that they exist in the sense of having themselves at the heart of their judgements, that is, whether or not other systems are self-conscious. Monkeys such as Imo would certainly seem to

exhibit signs of being self-conscious, and monkeys and apes are so close to mankind on the phylogenetic scale that I would feel uneasy about denying that they possess a self-conscious existence. Likewise for other mammals and even reptiles and birds. The task of attributing self-consciousness as a mental state becomes more difficult as the behaviour exhibited by the species in question becomes less and less human-like. Thus when we arrive at systems such as fish, insects and invertebrates it is more difficult to attribute self-consciousness because the behaviours these systems exhibit are so far from those that we usually associate with human self-conscious behaviour that we begin to doubt they really are self-conscious after all. It is not a problem to attribute consciousness to such systems because they are reasonably complex in their architecture, environment and behaviours, and on top of this they are capable of adapting to survive in their changing environment. They must be capable of perceiving their worlds, taking in, selecting and processing information and responding to it in the way that will enhance their chances of survival. That they can form beliefs is unlikely and for insects with short life spans completely unnecessary.

When we arrive at the cluster diagrams for systems that are not alive but which can exhibit human-like behaviour, for example, PDP's, we again find ourselves in a quandary about whether or not these systems ever could be, self-conscious. It is true that they exhibit learning behaviour, but that they can form beliefs from this newly learnt information remains at least, highly questionable and at most, extremely doubtful. They may be able to simulate a process of belief formation, but this is not forming beliefs in the human sense where all previous experiences, the present environment, the possibility of the future outcome and the individual's personality come into play. It is true that PDP's are adaptable, and in possession of an impressive array of capabilities, but on the scale of living systems they are still very simple, so it is extremely unlikely that they can form beliefs from the information they receive, process and store in data-banks as 'knowledge'. Whether or not they could ever form beliefs is still an open question.

Some sort of continuum of lower order mental states can certainly be seen to exist among simple systems, for example, thermostats, invertebrates, fish, video recorders, and progressively right up to complex systems such as, weasels, antelope, and human beings, for they can all process information and to some extent refer to former states. The continuum for storing knowledge or data can also be seen to exist in some of the more complex and adaptable systems, but it starts much higher up past the level of amoeba, televisions and thermostats. However, the continuum ceases when we reach higher-level mental states such as being able to form beliefs, being self-conscious, being able to attribute mental states to other systems and ascribe meaning to symbols. Human beings are capable of these things, and when it comes to forming beliefs and being self-conscious, we might find that other higher-order primates and perhaps even some of the other mammals are just as capable; but when it comes to the ascription of meaning and the formation of language, being social animals capable of attributing mental states by examining relevant behaviour and forming analogies with our own behaviour, then human beings are the most capable, adaptable and complex of all systems. Being the only system that we know to be capable of all these complex behaviours has enabled a fundamental distinction to be made between our mentality and the putative mentality of machines, and it is the interruption of the continuum that makes human beings distinct from all other systems.

Endnotes:

[1] My knowledge of my own mental states is also fraught with difficulty. For instance, I frequently come across work that I have written and I have completely forgotten that I ever knew anything about it. Or sometimes it is only when in an argument with someone that I realise my beliefs about an area of politics are particularly fervent. So I do not have anything like a complete knowledge of my own mental states. Thus it can be argued that the notion of "privileged access" is also a very dubious one. However my mental states are the only ones of which I can have direct first-hand knowledge so in that sense they are the only ones I can ever really know. The mental states of others I take on trust, experience and analogy with my own mental states.

[2] I say 'theoretically' for TM's are not used in the construction or interpretation of formal languages because an FSM or a Push Down Machine (PDM), that is capable of recognising context-free grammars, can carry out these tasks easily. Nor is the TM used for generating or interpreting natural language for even sentences with ambiguous meanings can be coped with using a TM with a tape restriction because it is sensitive to context.

[3] Once again we are back to Nagel's argument that it is not possible to know the mental states of another system for one cannot imagine what it would be like to be that other system.

[4] Wittgenstein, L (1958) *Philosophical Investigations*, Basil Blackwell, paragraph 71

[5] Ibid. paragraphs 66, 67 ff.,

[6] The phrase "cellular automata" was brought into being by Joseph von Neumann. He believed that fundamentally life was logical and the 'food' of this logic was information, so that theoretically for von Neumann there was no reason for a machine not to have life. "Cellular automata" are his mathematical formulations of this machine life.

[7] The "Game of Life" was developed by John Conway towards the end of the 1960s. His intention was to show that from a simple but random beginning a pattern and complexity would quickly emerge.

[8] Example relationships with "non-physical" entities can be seen all around and are fairly wide ranging, but two come to mind quickly, the first is the imaginary playfriends that children have, and the second is praying to a deity to help us in a time of crisis. Neither of these are physical entities of which tangible experience is possible.

[9] A noted exception to the 'social' rule is the orang utan for it spends most of its life alone only coming together with another orang utan for brief periods to mate. More is said about the solitary habit of the orang utan in section 6.5.1.

[10] Attenborough, D. (1979) *Life on Earth*, p.181-183, William Collins Sons & Co. Ltd

[11] See also section 3.7., chapter three.

[12] This diagram is based on the venn diagram tradition since it is about the relationships between sets of things. However, it should not be mistaken for an actual venn diagram because it is not about the logical relationships of set theory; it is simply about recognising where overlapping mental states are likely to occur.

[13] The complete set of points and *Mathematica's* lay-out is included in the Appendix 4.

[14] *Mathematica* is a brand name for a computer generated mathematics application.

[15] It is important to remember that the values attributed to the points have been based on where each system lies on the diagrams with regard to the chosen set of axes. Each diagram has been discussed at length with colleagues and friends so that my choice of positions for the clusters would not seem spurious. However, if a different set of axes had been chosen the clusters would have been arranged in a different way and if a different range of measurements had been chosen the values of the points would have changed. The outcome of this is that these diagrams are intended only for discussion for they have no strictly, empirical or mathematical basis; but in its favour it should be said that with the addition of the third dimension, and a grid upon which the points can be given values, a more informative representation is achieved from which it is easier to deduce the presence, or at least the likelihood, of a systems having mental states.

[16] Kelly, K. (1992) "Deep Evolution, The Emergence of Postdarwinism", *Whole Earth Review*, p.16

[17] Shakespeare, W. (1604) *Hamlet*, Act I, Scene V, lines 106 - 108(inc.). Penguin Books, 1969.

[18] "*Silicon Souls, How to design a functioning mind*", 18 May 1992. Taken from the seminar paper.

[19] Ibid. p.6

[20] Ibid. p.19

[21] Underlying dispositions of the personality are certainly not something new for they were current as early as Chaucer's time in mediaeval England. The four categories on which personalities were always based were *Phlegmatic* - water, *sanguine* - air, *choleric* - fire, and *melancholic*- earth. Of course, the divisions into the four elements went still further back to the time of the ancient Greeks.

[22] It is less likely to be whether one type of system, the organic, needs food and water whereas another type of system, the inorganic, needs silicon chips and electricity!

[23] "The great red-haired orang of Borneo and Sumatra is the heaviest tree-dweller in existence. A male may stand over one and a half metres tall, have arms with a spread of two and a half metres and weigh a massive 200 kilos." Attenborough, D. (1979) *Life on Earth*, p.285 - Animals such as these only come together with another of their kind in the mating season.

7. Conclusion

7.1. Introduction

Having now covered the groundwork I will begin to draw the threads of my argument together in this chapter by reiterating the conclusions that were derived in each of the previous chapters. Having completed this I shall move on to examine which characteristics we accept as being essentially human, and as far as we know belonging to no other system. To follow this up I shall examine the advantages that these characteristics offer the human system enabling it to take up, and so far remain in, a position of intellectual dominance within the world. Of these characteristics there are three that I will set out at some length.

The first is that the human system is the only system we know, as yet, to be capable of creating and arbitrarily assigning meaning to symbols. I shall argue against the proposal that symbols have an intrinsic meaning by virtue of the fact that they are symbols. Secondly, the human system is flexible enough to select the piece of information that is most important to it from its perceptual input. It has also the flexibility to form beliefs about the information, to influence the beliefs of others, to store the information for later use and so on. Implicit in this selection of the most relevant piece of information is the choice of what information is to be 'selectively' ignored. Related to this is the third characteristic, that the human system is capable of subjectively interpreting the pieces of information that it selects for particular attention. Human beings can see the world from their own point of view and are able to express their interpretation of their world using propositional attitude statements. The nature of mental states means that we do not, and perhaps cannot, know if any other system is capable of seeing the world in quite the same manner, and, if Wittgenstein is right, if another system could express itself using a language it would be impossible for us to understand what it was saying because its language would be about its relationship to its world. Each of these three characteristics will be discussed in much greater detail

before I move on to examine what sort of architectural, behavioural and environmental requirements are necessary for such a highly developed and complex cognitive system to exist and behave in the ways that it does.

This chapter will be brought to a close with some reflections on why it is impossible[1] for me to have direct knowledge of the internal states of any other system. I can only have awareness of my own subjective states and all other behaviour must be 'as-though' the system knows or 'as-though' it understands. Thus when we talk of the experience or ascription of mental states we realise that we occupy an asymptotic line[2] for although I have direct experience of my own states there are still problems with 'privileged access' because I cannot ever know all my mental states, and even though other human being would seem to display self-consciousness their self-consciousness is one of which I can never have any direct experience at all, nor they of mine. Thus, we are "destined to describe an asymptotic curve, which approaches but never reaches the limit"[3], the "limit" being our full understanding of self-consciousness, whether our own or that of other people. However, the reason we think that some human 'self-conscious' behaviour, such as the use of propositional attitude statements, is more likely to indicate a manifestation of self-consciousness in another person is due to the fact that we can imagine what it might be like to be the person behaving in that particular way. On the other hand the reason we are less likely to think of the machine or sea-cucumber as possessing self-consciousness is because it is impossible for us to ever fully imagine what it would be like to be a machine, a bat, an amoeba or a sea-cucumber and behaving in the way that they do. On top of this each of these systems occupies their own very limited environment, within which it is doubtful that they exercise subjective judgement.

Quite simply it is more realistic for me to guess at the mental states of other human beings, and that their 'as-though' behaviour is a reflection of their possessing mental states that are nearer in kind to the mental states that I, myself experience, because they are more like me than any other kind of system. Thus, all the mental states that are made manifest by other systems are similar in kind (to a degree that varies depending

upon the system) to mine, although there is no way that I can ever know their mental states as mine. Thus I can only ever be sure that I have self-consciousness, and that it is highly probable that other human beings have self-consciousness because we have so much in common, for example, our physiology, our social behaviours such as the use of language, physical mannerisms, and so on. That I have self-consciousness suggests that it is highly likely that other human beings have it too and that all of our conjectures about mentality in non-human systems will go on being hopelessly inaccurate because of the elusive and seemingly boundless nature of mental states.

7.2. Drawing the conclusions together - what has been achieved?

In the previous chapters I embarked on an attempt to answer the question, "When is it justifiable to say of a non-human system that it has mental states?". The search for justification was based on which mental states can be accepted as preceding and permitting particular actions in a variety of systems, both human and non-human, and I started by offering a critical review of some of the most relevant parts of the wealth of literature in the areas of philosophy, artificial intelligence, and cognitive psychology. This was, to some extent, 'scene setting' for it meant that some of the problems that are discussed and disputed in the theories of mental states and intentionality could be aired, explained and criticised whilst at the same time giving a clear indication of the direction that the thesis was going to take.

This account of mental states and intentionality then led up to a more detailed examination in chapter three of how we come to recognise the occurrence of different mental states and how we then go about ascribing them to other systems. These 'other systems' included other human beings, non-human animals, machines such as thermostats, televisions and PDP's and lower order organisms, such as amoebae and sea-cucumbers. I concluded that the ascription of mental states depends upon our apprehension of two things in the other system, (i) that it behaves in a consistently human-like way so that an analogy with our own behaviours and mental states is

possible, and (ii) that the other system is considered to be a complex enough system for it to be capable of possessing the level of mental states we are ascribing to it.

The actual business of ascribing mental states to another system can be carried out in two different ways. The first is by using language in the form of propositional attitude statements to describe X's behaviour, for example, that 'X believes that Y' or that 'X knows that Y'. When ascribing mental states to another human being the person to whom the states are ascribed can offer corroboration or denial of the ascription by herself using a propositional attitude statement to say 'Yes, I believe Z'. The second method of ascription is behavioural and it takes place, often without the awareness of the ascribing system, when the system doing the ascribing perceives that 'X' possesses a set of internal characteristics and then behaves in a particular way to that system; the 'poking and fiddling' approach.

Thus it became clear that the ascription of mental states is by no means a simple procedure for all of the criteria depend upon our subjective view of the world and the information we perceive, select and attend to from it. But if ascription is a difficult thing to do with any degree of certainty why do we do it at all? Well, one of the main reasons is that it is a useful predictive tool that facilitates interaction and communication between human beings and what we perceive to be other 'intelligent' systems. So that even if the system, whilst exhibiting signs of mentality, is still known to be inorganic, it is probably best, or at the very least useful, to behave towards it as one would towards a human being that is known to have a brain and a complex mental life.

With the ascription of mental states turning out to be such a complex process I examined three notions of complexity in relation to ascription in chapter four. The first one was that a system would have to have a fairly high degree of architectural or structural complexity for us to think that it could act in a way that is sufficiently 'human-like' and that would 'persuade' us that it ought to be ascribed mental states. Following on from this the second notion was of the complexity of the system's actions or behaviour, for simple behaviours would not inspire the ascription of high level mental states and the more complex the behaviour the more likely we are to ascribe

complex mental states. The third notion was that behaviour is a complex relation of architecture and environment, so that the internal design of the system and its environment afford the system a variety of capabilities, some of which will be more complex than others.

I concluded that a positive relationship exists between the overall complexity of the system, that is, its complexities of architecture, behaviour and interaction with the environment, and the system's capabilities to perform certain actions. If we then look at a computer its capabilities are constrained by a combination of its architecture, its program and the environment in which it is fixed. A similar state of affairs exists for the capabilities of non-human animals and it is only when we reach the higher-order animals that we notice that they possess an additional capability, that is, they have the adaptability to choose what information they will attend to in their environment and from this selection they can decide how to respond to the new information. This self-conscious element of behaviour is certainly one of the most complex aspects of the capabilities we know to be possessed by human beings.

Two of the other complex capabilities that human beings possess are (i) the ability to survive in a complex social environment, and (ii) the ability to create, use and adapt a shared language within that society. Thus as self-conscious, language-using systems human beings are capable of both thinking and acting in their own best interests but also, if the need arises, of subjugating those personal interests for the wider benefit and survival of their society.

From these conclusions it was possible to see that the system's complexity related, not only to the internal and external architecture of the system, but also to the system's degree of flexibility to respond to the wealth of continually changing stimuli that surrounds it in its environment. Human beings occupy the most enriched environment and they can do a lot. They have a high-level of awareness, they can select the most relevant information from their environment, understand it and follow this up with self-conscious judgements about it. They can also describe what information is irrelevant and ignore that. It is also possible for them to anticipate how other aspects of their

environment will be affected by their judgements, and to either change those judgements or justify them. Thus, it was concluded that human beings are very complex systems indeed with a huge repertoire of capabilities.

Looking at the possible ascription of mental states to both human and non-human systems led to an examination of the way in which the internal states of the different kinds of systems are described. Chomsky deals only with machine states and his example demonstrates that machine states are definable, limited and calculable. Dretske's hierarchy deals with mental states as they are applied to both machines and living systems and his example serves to highlight (although this is not really his intention) the indefinability and vagueness of mental states.

Both hierarchies are necessarily limited; Chomsky's because it goes no further than machine states and Dretske's because he does not manage to fulfil his commitment and complete the hierarchy by giving us a system that is capable of accomplishing first and second level intentionality but not third. Dretske fails because he tries to define and stratify things that cannot be examined in such a forced and contrived way. The difficulty of the indefinability of mental states means that the relationship between capabilities and complexity will be easier to locate in a machine than it will be in a living system. With mental states being vague and non-quantifiable it is going to be nigh on impossible to distinguish between any of the higher-level mental states. There are many examples where this is the case and one might cite the difficulty of drawing distinctions between 'love' and 'infatuation' and, of course, between 'believing' and 'knowing'.

There were other problems as well, for in a hierarchical stratification the positions of the systems are fixed and invariable so that anomalies such as a thermostat's being altogether more capable than a human being at detecting slight variations in temperature could not easily be shown. So most systems occupy rather contrived positions on a hierarchical model which suggests that a more adaptable model is necessary before a realistic picture of the relationship between the capabilities, architecture and environment of different systems can be shown. In

chapter six this idea was followed up and I offered examples of some of the many alternative ways in which this relationship can be shown.

The new way of showing this information incorporates the idea of cluster diagrams taken from taxonomy. At least three things emerged as being highly significant. Firstly, the choice of which axes should be used, that is, which criteria offer the most accurate way of defining and identifying the presence of mental states. For instance, that I chose to examine mental states in connection with the complexity of the system and its environment, its adaptability within that environment and the capabilities that arise from the degree to which the system is complex and adaptable, undoubtedly meant that other possible information became peripheral or was even excluded altogether. So the criteria by which we choose to define a set of things are very important since even members of the same family can be omitted if a slightly different definition is employed. Secondly, the sort of representation offered in a hierarchical construction is only two dimensional and even in the less constrained cluster diagram a two dimensional representation is extremely limited. A limitation of this kind means that the quantity and accuracy of the information being presented is always going to be adversely affected. The only way to overcome this is to increase the number of dimensions, and perhaps even the size of the diagram, so that more information can be included and the accuracy of that information is enhanced by the introduction of another set of defining criteria. The third, and perhaps most significant conclusion, is that it is now possible to examine vague concepts such as mental states by using cluster diagrams, and because it is already possible to examine the determinate states of a machine, it is now possible to draw up a comparison between the complexity, adaptability and capabilities of human and non-human systems so that machine states can be discussed alongside mental states. So we can now discuss the possible manifestation of mental states in non-human systems.

In the sections that follow I shall respond to the original question of how justifiable it is to ascribe mental states to non-human systems.

7.3. The advantages of the human system

There is nothing very new in claiming that human beings are more complex, more adaptable and more capable than any other mechanical or carbon based system (though it is recognised that these are not mutually exclusive), however, an examination of what are the necessary prerequisites that have brought about this state of affairs of high-level complexity might well yield some interesting and significant results. One of the fundamental requirements is that the human system must possess a rich and varied mental life with the capacity to form mental states as simple as those required to process incoming information and as complex as those through which it is possible to form beliefs about their worlds. What is also apparent in human beings is their capacity to create and use a shared language to discuss their beliefs and to describe an, otherwise unreachable, inner life of feelings and intentions. I shall now look at this ability in greater detail.

7.3.1. The creation and ascription of meaning to symbols

Within the area of the ascription of meaning there is an important division; it is between, on the one hand, the creation of a symbol and the subsequent ascription of a meaning to it (when this is carried out for sets of symbols and we then create strings of these symbols we have the logical beginnings of the development of a symbolic language[4]) and, on the other hand, the ascription of meaning to a non-symbolic state of affairs, or the interpretation we offer for a specific behaviour or type of behaviour that we believe is a symptom of a particular state of mind. This second type of ascription, of mentality on the basis of perceived behaviour, has been quite roundly dealt with in chapter three so little more will be said about it here. Instead I will concentrate on the first kind of ascription, that of assigning a meaning to a symbol or set of symbols.

The written word or symbol was created for the purpose of communicating between human beings in a more effective way, and by their creation and use the form of communication changed so that it could be made through time from one generation to the next, no longer relying merely on the vagaries of the spoken word that could be

misheard or misinterpreted. There can be little doubt that the spoken word preceded the written word and it was only with a demand for more information that the need arose for the spoken word to be transferred to a written form and this need was accompanied by a need for symbols.[5] Then came the ascription of meaning to those symbols for a symbol cannot be said to have a meaning by virtue of its being a symbol, which is to say, symbols have no intrinsic meaning. In fact a symbol only becomes a symbol when it is symbolic of something and that meaning is then designated arbitrarily to the symbol by its creator.

The ascription of meaning to a symbol can then be due to one of at least three possibilities. Firstly, that there is some aspect of our representational system and our ability to behave intentionally that is intrinsic to us as human beings; Searle's position. Secondly, that there is no intrinsic meaning present in any symbol, but that our ability to systematically interpret symbols and symbol strings is intrinsic. This suggests that a symbol's meaning is ultimately embedded in some non-symbolic representation; Harnad's position. And thirdly, that the meaning we ascribe to any symbol depends upon the interaction we have with our environment, so that my language, the meanings that are ascribed to my words and my environment are inextricably linked. I maintain that this last alternative is the one that is most probable for it does not rely upon the meaning of our symbols or their formation and interpretation being in any way intrinsic to us.[6]

Rosenschein agrees with this for although machines are quite obviously capable of manipulating symbols only the programmer is capable of assigning any meaning to the symbols.[7] And, for Rosenschein, if the creator of the program wishes to assign a different interpretation to the same symbols the machine will have a different sort of knowledge.[8] So the process of meaning ascription is arbitrary as long as each meaning is consistent with the others. Thus the meaning of the symbols is very important for the description of the machine's internal state, or what might be described as 'mental state' in Rosenschein's more limited and purely logical sense. Indeed Rosenschein recognised the importance of the environment and created the Situated Automata

approach where the state of the machine is a direct result of the limited interaction it has with its environment. Then when we ask the question "What does the system 'know'?" we can answer that it 'knows' the information that has been instantiated by the programmer plus the things that it has reacted to in its environment, and even though the things it reacts to have been dictated to it by its program there is a sense (again a very limited sense), that because it has been affected by these things it 'knows' or is 'aware' of their existence.

Another way of saying this is that a symbol only becomes a symbol when it is about some object or state of affairs in the world. Searle would agree with this but he would add that as the ascription of meaning must come from outside the symbol system, and as only human beings have yet proved capable of such ascription, human beings must have some intrinsic characteristic. This characteristic, he would go on to argue, is their own mental content or internal semantics. So intentional states must have their own intrinsic or self-attributed content and it is this element that he claims is missing from any computational simulation of human mental states. As proof he adds that an application in a machine will continue to run regardless of the fact that the machine understands nothing of the symbols or program that has been instantiated in it.

From the work that has been carried out in this thesis I would not wish to dispute Searle's claim that the machine does not understand the program that it has running, but I would say the machines lack of understanding is due to two things. Firstly, the fact that we have only a limited notion of what understanding is and how any manifestation of the mental state of 'understanding' should be discriminated from another mental state, for example 'recognition', and identified is not at all clear. And, secondly, that the attribution of vague mental states to things that operate on the basis of states that are by nature discrete and definite, is simply wrong-headed and perhaps even a matter of some vanity.[9]

No, my argument with Searle is with his idea that human beings somehow have intrinsic intentionality, that is, that the human representational system has its meaning intrinsic to it. Harnad also challenges Searle on this point but I believe Harnad's

argument is brought down by the fact that he still argues that because symbols are interpreted systematically by a machine, or for example, the individual in the Chinese Room, the form that the interpretation takes must be intrinsic to the system.

Harnad argues, and I agree, that the symbols in a formal symbol system only have meaning when they stand for things in the world. Meaning of this sort cannot be intrinsic to the system since any meaning is going to be based on what the symbols mean for us. So the interpretation depends on the fact that "the symbols have meaning for *us*, in exactly the same way that the meanings of the symbols in a book are not intrinsic, but derive from the meanings in our heads".[10] The symbols in any book will only have a meaning when we know the language they are written in and we can attribute a meaning to them. Harnad describes this as a *"merry-go-round"* because the attribution of meaning to one symbol always depends on another symbol, and for him the only way to get off this *"merry-go-round"* is to ultimately ground the meaning of a symbol in some non-symbolic representation. But these non-symbolic representations will themselves have to have an intrinsic meaning if we are ever to get off the *"merry-go-round"* . The example given by Harnad is of a 'zebra', which he argues we know to be the combination of essential and necessarily unvarying features of the two symbols, 'horse' and 'stripes'. The naming, he says, is immaterial once we have the composite meaning, so that the combination of 'horse' and 'stripes', the 'zebra', could have any name but it would always have the same composite meaning. In fact Harnad would probably want to argue that it is the 'naming' aspect and nothing else that is influenced by society and our environment, while I would argue that it is the semantics of our symbols and symbol systems that are grounded in inter-personal exchanges.

Harnad goes one step further in his argument saying that the individual who has the meanings has them in 'isolation' when he or she is combining 'horse' and 'stripes' to form the new concept of 'zebra', but this seems to be very curious notion for it seems impossible to ask what is going on in the mind of the individual when he or she has meanings in isolation from any social or linguistic interaction.[11] That an individual exists to have these 'isolated' meanings is surely indicative of an environment in which

he or she must be existing and in which his or her thoughts can be had; thus, the only feasible solution to the problem seems to be that there can be no difference between what Harnad describes as 'intrinsic meaning' and what would elsewhere be described as the individual giving what he or she perceives a 'subjective meaning' or 'interpretation'.

So Searle says that the meaning of any symbol is intrinsic to us as part of our representational system, and Harnad says that symbols come to have a meaning once they are used compositionally in meaningful syntactic ways, and that there is a difference between intrinsic meaning and extrinsic meaning, for the former is in our heads and the latter is attributed from outside. This is a very difficult idea for Harnad maintains that we go through the perceptual phases of recognition, discrimination and identification, after which we reach a representational level at which categorisation takes place where a meaning, but not necessarily a name, is attributed to them. Finally Harnad argues that the meaning of a symbol is grounded in non-symbolic representations that are formed into meaningful and syntactical strings of symbols that can then be interpreted systematically and it is this interpretative ability which is intrinsic to the system.

But what can be the nature of Harnad's non-symbolic representations? It is not at all clear for they might be like Wittgenstein's 'objects' or 'things' in the *Tractatus* that can be shown but not said,[12] or even like his 'family resemblances' in the *Philosophical Investigations*, that can be described and examples given to show the inter-relations but nowhere can the overlap between two things of similar type be explicitly stated. Another possibility for non-symbolic representations is that they might be like the Platonic 'Forms', where for example, the *essence* 'horseness' or 'tableness' can be sought. The nature of a non-symbolic representation is certainly not immediately obvious which makes Harnad's brave leap from symbolic representations to non-symbolic forms of representation seem altogether odd and perhaps even futile.

However, there is one redeeming feature in Harnad's theory and that is that it appeals to the individual's world so that any assignment of meaning to a symbol will be

made on the basis of the referent of the symbol already existing in some form in the world of the person doing the ascribing. Searle also says this but then carries on to say that the meanings we ascribe are an intrinsic part of our representational systems and it is here particularly that I have a grievance with his theory, for I would argue that the environment, or context, in which the human being lives and breathes is all important for making meaningful ascriptions to the new symbols that he or she has created, and also for the assignment of new meanings to old symbols. If the meanings were implicit in the human representational system the interpretations of every situation ever encountered would be identical for every culture and plainly this is not so. Our languages are different, our cultures are different, for example, to the American Indians the natural world talks but to the Americans in New York city it is money that talks. The same things have different degrees of significance to different cultures which suggests that the social and cultural environment of the individual is the most important part of the ascription of meaning to symbols, and that there are such vast degrees of variation in meaning for the same things casts doubt upon there being such a thing as an intrinsic representational system.

For a symbol to be symbolic it has to represent some state of affairs that exists, either intentionally or actually, in the world of the person doing the ascribing else the symbol could not be said to be truly 'symbolic'. However, it is true that this 'something' can be either physical or abstract in nature, since the environment of the individual includes the world of *phenomena* that exists externally as a world of 'appearances' and the world of *noumena* that exist as a form thought or intellect, the 'things in themselves'.[13] Examples of phenomena in our world are prolific, for example, trees, biscuits, cats and so on; but giving examples of noumena is not so easy, perhaps, a number in pure mathematics[14], or the feeling of anxiety or envy.

So because we live in a shared world in which there are things that we want to talk about a need for communication first arises. Then by representing these things symbolically and ascribing a meaning to those symbols the information that we want to pass on can be conveyed in a sort of shorthand. As in chapter four, section 4.2.2.,

paragraph 10, when I said that no behaviour can take place in a vacuum, so too our ascription of meaning to symbols is a form of linguistic behaviour and it cannot be exercised without the presence of an environment. Thus our interaction with our environment sets limits to the creation of language and the ascription of meaning to our words, for it is not possible to talk of things of which we have no concepts for we would have no words and hence no way of describing them. It is certainly true that we devise mythological characters that do not exist in our world but such beasts are made up of a collection of the parts of other beasts that do exist (or at least have at one time existed) and of which we have logical and linguistic conceptions. For example, the *griffin* is a conjunction of a lion's body and an eagle's beak. Indeed most of the examples of beasts in the genre of Science Fiction bear this out for they rarely look like anything other than oversized beetles or other hideously enlarged insects!

There is no need for us to rely upon the meaning of our symbols or how they are formed and interpreted being intrinsic to us for we have seen in the previous chapters how complex, adaptable and altogether capable the human system is in comparison with all other systems and species, and it is because the human being possesses such capabilities that it is able to develop by seeing how something could be done better, creating where there is a need, utilising all facilities and surviving against all odds. The development of language is only one example of the great wealth of human capabilities but it is surely one of the most impressive and distinctive. It is not that I doubt that there may be something in the structure of human DNA that gives us the propensity to understand how symbols can be made to stand for things and how those symbols should then be manipulated in language so that their meaning can be obtained, but I do maintain that none of the creation of symbols, the ascription of meaning to them or their manipulation could have been carried out without the existence of a social environment in which these symbols have a use. After all it is only within a social environment that any linguistic communication would even be necessary.[15]

With language I am able to describe the feelings I have and express my intentionality in a definite and determined way that can be understood by other people

who share the form of language that I use. I can express myself and advise other human beings of my intentions, I can also understand more readily how they feel about a state of affairs since we are in possession of a shared method of communication. However, when this method fails, as for example, when I am in the company of someone who comes from a different culture and speaks a language that is foreign to me, I have to rely upon their facial and bodily expressions to convey to me their wishes, beliefs, and so on. Without the use of a shared language I am more likely to misunderstand the other person and ascribe to them states of mind that they do not in fact possess. So we can see that a shared background or environment and the use of a shared language that has been derived from the same cultural environment is the most reliable method of communication with another human being. Other non-human systems communicate with one another in a much less sophisticated manner than do human beings who have a shared language and background, for other systems do not employ anything as complicated as symbolic representations or notation to convey their meaning.

The ability to create symbols, to arbitrarily assign meaning to those symbols, to then form those symbols into strings that are syntactically correct and semantically interesting, and finally to use those strings with the intention of communicating information to another human being is just one of the advantages that human beings have over any other system. That there is a need for language at all must surely be a simple reflection of the complex nature of the human beings environment, architecture and subsequent behaviour.

In the next section I will examine a second advantage that human beings possess, that of being able to select which is the most relevant piece of incoming information for them, understanding that information and using it in ever changing environment, whilst also being able to selectively ignore that information that is not of interest or use. Because we face problems experiencing the states of another system we cannot say for sure that any other system understands its environment to the same level of understanding that human beings understand theirs, thus it is impossible to say whether

or not other systems 'know' or 'believe' the events and states of affairs that are taking place in their worlds. The abilities to select information, understand it, elevate it to the level of knowledge from where beliefs can be formed about it will be assumed to be advantages that belong in full solely to the human system.

7.3.2. The ability to select information for attention

Being capable of the creation, ascription, formation and employment of symbols and strings of symbols is certainly one of the most impressive of all human abilities and one that is not shared by any other non-human system. In this section I will look specifically at what other human capabilities can be inferred from this sophisticated use of language.

Any system ranging from the low-level thermostats and amoebae to the high-level organisms such as mammals, monkeys and human beings are capable of processing information. They are capable of responding to specific things in their environment but the flexibility to choose which things they will respond to varies quite significantly between species. For example, a thermostat can only react to one thing in its environment, and that one thing is any change in the temperature of the room it is monitoring. If it detects any variation it can act accordingly by turning the heating on or off. In circumstances where there is no change in the room temperature it will do nothing. It is capable of no more than this, and is only capable of less than this if it is broken or disconnected. Human beings, and perhaps a great many of the higher-order systems, can respond to any of the things that they perceive in their environment. They have the flexibility to select what is of most interest and attend to it whilst perhaps retaining some of the other perceptual input as stored information that can be attended to later when it might be more relevant. Thus it is that the flexibility to select that which will be attended to and that which will be ignored is one of the characteristics that sets human beings, other higher order primates and some of the other more capable systems apart from the simpler ones that have a fixed environment in which they can only attend to a specific set of things.

However, it is also true to say that we cannot be certain that any other non-human system can select and ignore information in just the way that human beings do, but having watched my cat playing by actively seeking a ball that it wants to play with or wait in ambush of another cat, or watching a group of chimpanzees taking turns scrutinising a leaf and each of them looking serious as they do so and then simply tossing it aside, it would seem that there is much in their environment that they choose to ignore and at particular times there are things that are given special attention at the expense of everything else. Naturally, there are times in the existence of even the most complex and capable species when they are not given any choice about what they can respond to, such as being in imminent danger and needing to get away, but at times such as these, without the element of choice, the response becomes an intuitive or 'gut' reaction and not something about which there can be any deliberation at all.[16]

In the main the more complex, more adaptable and thus more capable systems also seem to possess the flexibility to choose what things they will respond to in their environment. Therefore the nature of the system's environment[17] is also of great importance. The thermostats environment is fixed and extremely limited. Although it can perform a greater range of functions the video recorder still has a limited, fixed environment. It is only when we reach the level of PDP's that machines become slightly more capable because they have an environment with fewer limitations and more possibilities. As a result the PDP is more flexible but there is still an order in which it will perform whatever functions it has to carry out, for after all it is dictated to by a binary machine code. Indeed even when we examine non-human animals it is difficult to establish areas where there is much flexibility in their choice of what to respond to and what activity to carry out next. Animals, such as, protozoa and sea-cucumbers have limited environments and in fact it is only when we reach the level of the high level mammals that there seems to be much flexibility at all. By and large the behaviour of non-human animals is dictated by their physiological needs and it is only when we observe an animal playing, which does not seem to be a goal-directed

behaviour, that the element of choice seems to enter into their repertoire of behaviours.[18]

The environment that human beings occupy is vast and unlimited for not only can they select the things in their world that they will attend to, but they can also think about a state of affairs when it is not immediately present, that is, in reflection, or they can speculate about the existence of an omnipotent creator, or aspects of natural philosophy plus much, much more. So the choices made by any human being will include things that are within, but also without, its perceptual domain. Understanding the amount of flexibility that human beings possess is made possible by their use of many shared languages that enable each person to know to at least some degree of certainty the form and content of another human beings mental life and mental states. Human beings can share their experiences and thoughts through discussion and similarities between someone's professed mental states and their observed behaviour can be implicitly drawn up as a parallel or analogy with my own mental states and my own behaviour. It is only when the parallel cannot be drawn, for instance, when someone claims to be a caring person and then berates people claiming social assistance for being lazy, that we begin to question the integrity, or even the sanity, of that person.

So when it comes to selecting what things will be attended to and what things will be ignored there first needs to be an environment in which an alternative is possible. For the thermostat and video recorder this is not the case. For a PDP the environment is enriched for it can learn from the information it receives but it still follows a course of events that is dictated by its design and program specification. In the non-human animal world the environments are, by and large, richer with more diversity but the system still attends to those things that it needs to satisfy a physiological need. It is only when we reach the higher order mammals that the element of choice begins to play a more central rôle in their behaviours. However, it is still only when we talk of human behaviour that we can say for definite that here is a system that chooses the things in its environment to which it will respond. Human beings frequently behave in ways that are not dictated by any physiological, and certainly not logical, necessity, indeed what any human

being chooses to respond to is more often a matter of that individual's subjectivity. For example, I choose what I will read in the newspaper because I know what is of interest to me, in a similar way when I look out of the window I look in the direction from where I expect someone will arrive and only if something dramatic enough to overrule my choice occurs will I attend to something else. The choice that a human being makes is not necessarily dictated by physiology nor logic, and is often a the result of their own subjective nature that they decide to follow one course of action rather than another. It is this, the individual's, element of subjective interpretation that I will explore in the following section.

7.3.3. Subjective interpretation

The choice of what actions we, as human beings, pursue is largely our own decision, a matter of our own subjective choice. The decision ceases to be our own on occasions either when someone else tells us what we should do, for example, the person in charge at work, or when we are driven by a physiological need to find food or water or the urge to satisfy a sexual need.[19] Perhaps what is more interesting is that of the information we choose to attend to we can offer a subjective interpretation and that this interpretation will be unique to each of us as an individual. The uniqueness criterion is fundamentally due to the fact that no one else can ever have had my experiences and that all of the experiences in my history are personal to me.

It is simply that I see things from my own point of view, no matter how open minded I am![20] The 'I' is present in all of my judgements because everything I decide to do will affect my life in some way. I do not live passively in my environment for everything that I choose to attend to and even the things I choose to ignore have a meaning for me, and I cannot fail to ascribe one to them. In the case of the things I choose to ignore the meaning is likely to be cursorily applied because I know the thing has little or no relevance to me. The things I choose to attend to might have no meaning for me as yet for I may know nothing about them and be actively seeking more

information by attending to them, but they are things that I believe will have relevance for me so already they have a meaning in the sense of being 'meaningful' to me.

Just as human beings are aware of their informational input so all non-human systems must be too else how could they survive. Even a machine is aware in a very limited sense of input from its environment, if it was not it would simply be a collection of bits with no function. The difference between human beings and other systems is that they are self-consciously aware of the information they receive through their senses and they interpret it in relation to themselves. That I am capable of seeing myself at the centre of the judgements that I make about the information that I receive, and that I can describe this relationship between myself and my environment using propositional attitude statements is an indication of the great distinction between me, as a human being, and all other non-human systems.

The immensity and variation in the human environment is again significant for not only does it offer us a great wealth of information from which we can choose that which we will attend to and that which we will ignore, but it also means that any behaviour we exhibit or we observe can have any number of interpretations, for the interpretation we assign to any behaviour is heavily influenced by the context in which that behaviour is performed. For example, were I to see someone sitting alone and crying I might think that the person is obviously very sad and in need of consolation, whereas, were I to see the same person exhibiting the same behaviour but this time in a theatre, watching a production of Sheridan's "The Rivals", I would be more likely to think that they are enjoying the play and that their tears are tears of joy. So three things that are significant are, (i) the vastness of the human environment, (ii) the infinite number of possible human behaviours and (iii) our ability to interpret everything subjectively.

Indeed with our subjective interpretation of events being able to stretch into the realm of abstract concepts and ideas means that I can ask myself questions about my existence, the infinite nature of the universe, whether there really is an after-life and so on. At first it may seem an odd idea but the environment in which abstract thoughts are

aired and considered is also very important for their interpretation. For example, if I attend a lecture on "The mind as a control system" given by Aaron Sloman as part of the proceedings of a Royal Institute of Philosophy Conference, I am more likely to listen to what he says even though I may disagree whole-heartedly with what he is saying; however, if someone stands up on a soapbox at Hyde Park Corner and tries to tell me that it is possible to build a machine that, with the addition of a "chemical soup", could respond to information and have emotions just like a human being I would be more inclined to think of that person as over-optimistic, if not plain crazy.

The possibility of abstract thought in other systems cannot be ruled out on the basis that because I do not share a language with any other system I cannot ask them, and they cannot tell me, if they think of things other than the physical information they perceive from their environment. Chimpanzees do exhibit a marvellous curiosity about their worlds and with their being so close to us genetically it might be that they too wonder about their existence. I would be loathe to rule this out completely. But it may be idle optimism on my behalf for we have no evidence that they do, nor even that it would seem interesting for them to do so.

So here we have at least three things that separate the human system from all other non-human systems; firstly, the creation of symbols and ascription of meaning to those symbols with the subsequent use of a shared language that has shared meanings; secondly, the flexibility to select the piece of information that is most relevant to the individual at any one particular time; and thirdly, the ability to interpret the information that has been selected in the individual's own subjective manner. The human system is without doubt one that is capable of complex cognition and I shall now look at what sort of architectural, behavioural and environmental requirements are necessary for cognition of this sort to be possible before rounding off with my response to the question posed at the beginning of the thesis.

7.4. The requirements for complex cognition

What characteristics would a system need to possess for it to be capable of complex cognition such as that carried out by human beings; well for a start it would need to be conscious of itself in relation to its world, so initially it needs, at least, to be self-conscious. A second characteristic is that it needs to be capable of selecting the most appropriate piece of information from the wealth of incoming stimuli that are bombarding its perceptual field at every moment of the day. On top of all of this the system needs to be capable of subjectively interpreting the selected pieces of incoming information, and this requires a great deal of flexibility from the system.

For the interpretation to take place the system needs to be capable of creating symbols and ascribing a meaning to them. This meaning has to be fixed in the sense that it can be shared with other users of the same symbol system without there being any loss or substantial variation in the interpretation of the symbols. This last requirement is arguably the most important for it is only through the possession of it that any system would be at all capable of expressing its self-conscious capabilities, the subjectivity of its judgements, and discussing information with other like systems and from those discussions and other incoming perceptual information forming beliefs about its world that will enable it to adapt and survive.

As a human system I have these capabilities with which I can recognise that my thoughts and experiences are my own. I can also speak about my experiences with other human beings who use the same language that I do. It is even possible for me to understand, to a limited extent, the actions of another human being with whom I do not share the use of a language and this is more strongly suggestive of the basic commonalities between my mental states and those of another human being. I can understand myself and others like me, and I can offer an interpretation of the behaviour of other types of system.

So for another system to be capable of complex cognition it needs to be capable of understanding the information that it receives and forming knowledge or belief states

from that understanding. As mental states are, by nature vague it is impossible, so far in our knowledge of the world, to say exactly what criteria would have to be fulfilled for us to recognise the form that understanding behaviour would take in a non-human system, or to differentiate between a 'knowing' state and a 'believing' state. Indeed as we have seen this latter example is difficult enough to settle in the case of human beings where we are in possession of a great many more of the facts.

7.5. Conclusion - so when is it justifiable to say of a non-human system that it has mental states?

Throughout this thesis we have confronted a great many difficulties involving mental states for they are vague entities that make it difficult for us to recognise their presence, differentiate between them and identify them absolutely. Therefore, it is difficult to ascribe them and differentiation of them is only possible on the basis of perceived complexity, adaptability and capability. It is because mental states cannot be identified and differentiated in the same way that machine states can, that it is not feasible to try and show them in the form of stratified hierarchies. In the cluster diagrams of chapter six it was possible to show that the human system is the only system that is capable of occupying a state of 'full blown' self-conscious awareness, and that although lots of other systems can occupy states of varying levels of complexity, none, but the human, language using, system is capable of the full gamut of known mental states.

There can be no doubt that it is a useful practice to ascribe mental states to both human and non-human systems for, as Dennett has argued, it allows us to more readily predict their behaviour and thus ourselves behave in accordance with what we anticipate they will do. Therefore, in the sense of being a useful thing to do the attribution of mental states to non-human systems might be argued as justifiable. However, if we stand back and ask is it justifiable, in the sense of can we say with any degree of certainty that these other types of system have mental states that are identical in kind to those that we ascribe to other human beings, then, after a lot of deliberation, I would have to come down on the side that maintains that it is not.

So it is that a distinction has been drawn up between two senses of 'justifiable', the first is having a (good) reason for doing 'x', and the second is being able to show that 'x' is conclusively the case. Each of these senses has its own precise application and use and it is in this area that our problems have been seen to arise, for often the language we use to describe and ascribe mental states is used thoughtlessly resulting in the misappropriation of mental state terms.

Even when dealing with human beings the ascription of mental states is fraught with difficulty. Eventually we have to decide which is the most sensible way to progress in our ascription and that is to first look at the other person's behaviour, compare it with my own behaviour and the mental states I would have that would accompany it. So it is simply a matter of optimistic analogy that is encouraged by the fact that other human beings look like me, talk like me, and act like me, so why should they not think like me as well. Any interactions I have had with other human beings have always been on a basis of ascribing meaning to their actions and so far this has been successful and there is no reason to think that my ascription will not also continue to be so. Thus it would seem realistic to assume that there are a great many commonalities that exist between my mental states and those of other human beings and that my certainty about the ascription of mental states to other human systems is indeed vindicated.

However, the character or nature of the mental states possessed by other non-human animals is not so easy to identify and pin down. We compare their behaviour with our own behaviour and the mental states that would accompany our behaviour, then we attribute to them these mental states and this may, quite simply, be completely mistaken. But it is all that we have to go on for we can not locate a mental state and analyse it in the way that a brain state can be isolated and examined. A brain state is necessarily something physiological and mental states are not, at least as far as we presently know.[21] Until we can share a language with another living system our theories about their mental life can be nothing but conjecture, and there is no certainty in conjecture.

When it comes to machines, such as personal computers and PDP's, we are more inclined to say of them that "they know what the symbol means" or that "they understand the task they are carrying out", but there are two reasons for this; firstly, they carry out tasks that human beings would otherwise have to do and they do so quickly and efficiently; and secondly, the usual type of interaction we have with a computer is carried out using a language that we understand. The computer gives us information in a language that we recognise and use so the computer and I seem to possess a shared language, and this is very misleading for it encourages us to attribute to the machine all of the mental states, and at least some of the capabilities, that we would otherwise only attribute to another human being with whom we can carry on this advanced level of communication. If we look more closely at the machine we quickly realise that any of its seemingly well-versed interactions are simply the product of human programming labour and that the machine does not understand us or the interaction we are engaged in after all.

To use the language of the mental, that is Fodor's 'mentalese', to ascribe mental states to machines is to over extend its use and it is done, not because we fail to understand the nature of machine states or that we fail to see that they, machine states, are different in kind to mental states, but rather that we understand so little about mental states that we are prepared to proffer their manifestation in even the most unlikely places in an effort to understand them more fully. Furthermore, it may be that we ultimately discover that this earnest ascription of mental states to non-mental systems has only furthered a misunderstanding of the nature of mentality and mental life.

Thus, I can only conclude that the ascription of mental states to machines may be a useful exercise that allows us to predict the outcome of any interaction that we may have with them, but really what we are doing is comparing two systems that are, by nature, more dissimilar than we usually seem prepared to admit. For the attribution of vague and non-quantifiable mental states to a system that has internal states that are by nature tangible, definable and measurable, seems just too much like anthropomorphism and the desire to play at being some sort of omniscient being that is capable of creating

things in our own likeness. If this 'likeness' includes our faults as well, and with the capacity for intelligence it may well include a desire for competition and dominance, then we enter into questions of the morality of bringing such a *being* into existence. But this is another question that could be raised in another thesis.

Endnotes:

[1] I am here assuming that ESP has not yet been shown to work, but I am not maintaining that it would be impossible for it, or some form of new technology, to allow me to gain direct access to the mental states of another system, and vice versa.

[2] See also section 3.6.1., chapter 3.

[3] Miller, J. (1992) 'Trouble in Mind', *Scientific American*, Special Issue - September 1992, pg.132

[4] A "symbolic" language as opposed to a non-verbal form of language such as, "body" language and facial expressions both of which can be as expressive, perhaps even more so, than an ordinary verbal form of communication.

[5] I do not wish to argue that this is the only reason that written language developed for many other factors were also influential. Indeed one well known example is that of people wanting to pass down stories of great adventures, or even of cautionary tales, to the generations that were to follow and the easiest and most lasting way to do this was through the written word.

[6] Of course, there are other possibilities such as Plato's theory that before our birth we have all possible knowledge which the shock of birth makes us forget and the rest of our lives is spent remembering things rather than having to learn everything from the beginning each time.

[7] See Chapter Two, section 2.7.1.1. and following.

[8] It should be remembered that 'knowledge' in Rosenschein's sense is limited to a logical encoding in a formal language. See chapter 2, section 2.7.1.2.

[9] The creation of images in our own likeness, that is, with human-like intelligence, is very much like the Christian view of God creating man in his own image.

[10] Harnad, S. (1990) The Symbol Grounding Problem, p.339, Physica D 42

[11] Ibid. footnote 2, pg.336.

[12] Wittgenstein, L. (1961) *Tractatus Logico-Philosophicus*, 4.1212 and 5.62. Routledge & Kegan Paul

[13] Kant, I (1787) *The Critique of Pure Reason*, p. 257 - 275 (incl.). Translated by Norman Kemp Smith (1929).

[14] Numbers in the sense of pure mathematics are not numbers in the Fregean sense, because he describes numbers as 'objects' for any number can be the reference of a singular term. This description means that, for Frege, numbers are no longer abstract terms but things that can be referred to and discussed in much the same way as we would discuss a horse or the score of an aria.

[15] See chapter two, section 2.5.1.2., for how a 'private language' is of no use for there could be no shared meaning and no communication. Also for a lengthier exposition read Wittgenstein, L (1958) *Philosophical Investigations*, and in particular paragraph 293.

[16] In the previous example of, the thermostat, I used the word 'react' instead of the word 'response', therefore, I would propose that a reaction is an action that has the element of choice taken out.

[17] In cognitive science the system's environment is described as its 'perceived domain', for the limit of any system's world is distinguished by what it can and what it cannot perceive. This notion of 'perceived domain' could be questioned in the case of the human species for they are also capable of abstract thought and creating concepts of things that exist outside their world, Kantian 'noumena' or 'things-in-themselves'. (See also section 7.3.3.)

[18] It has been proposed that playful behaviour in animals is goal-directed because it acts as practice for future confrontations with both prey and predator. I do not wish to argue with this view except to say that playing is a much less direct method of goal attainment than seeking water or food. The immediate goal of play is to expend energy and release tension, the indirect goal is to be prepared when a predator appears or there is prey in sight.

[19] I am aware of the moral tensions involved with the addition of the physiological need to fulfil a sexual drive, or of eating taboos, but here it is used merely as an example of a physiological drive and nothing more.

[20] See also section 4.3.4.1., chapter four for a full account of the notion of subjectivity.

[21] The nature of mental states is something that could be investigated in further work on this area. For example, if non-human animals have nothing but physiological drives it would seem most likely that the mental states that accompany them would be what I have described as brain states that can be isolated and analysed.

Appendix 1

Mediaeval Aristotelianism

A great deal of the work of the Mediaeval Aristotelianists, through the thirteenth and fourteenth centuries, was concerned with how the mind takes in information from the world and is then capable of processing and understanding it. Theology was given a scientific use, as a form of knowing; and as with the notion of God being something not phenomenal yet something toward which we could direct our thoughts and propositions, there was the origination of the notion of the intentional inexistence of objects. It was no longer necessary for every object of a statement to be a physically existent thing. Out of theology had grown a system within which it was possible to make belief statements about abstract entities.

Aquinas 1225-1274

The work of Thomas Aquinas is a fine example of the work of Mediaeval Aristotelianism for he tried to draw together the notions of theology and reason. He attempted to combine Aristotle's work on philosophy and logic with Christian doctrine and western ways of life. In *Summa Theologiae* he introduces the idea of reason and revelation as a means to knowing God. He offers five reasons for God which include both reason and revelation, and how through the human intellect it is possible to conceive of God.

If we regard the essence or soul as something that animates the physical body then all living things have a soul; vegetative in plants which is responsible for nourishment and growth, sensitive in animals because they are capable of sensation and rational in human beings since they are capable of rational activity. We are aware of material things external to us so we can not be entirely material ourselves. The human soul is capable of the sensitive and vegetative soul activities in its being able to understand

concepts and reflect upon logic, mathematics, metaphysics and God. Rational activities have no bodily counterpart. (The physical counterpart may be the rule base upon which computer programs are modelled or even neurophysiological changes in brain states that may one day be mappable.) For Aquinas this meant that there was an aspect of the human 'being' that did not have a comparable bodily activity and this would have to be immaterial and capable of surviving after the physical death of the body.

This rational part of the body contains the intellect and the will. The former is the power through which we attain knowledge, and the latter is the power through which we make choices. A well tuned intellect will be able to say what is good and what is not good. The will can only act in accordance with the intellect and being able to choose is the means to achieving that good. It is not the will as a means to an end that we are interested in, but rather that Aquinas considered that human beings were able to act freely.

Essentially a free act is one done out of a combination of reason and will, and the fact that we have a liberty in our choices is dependent upon the type of knowledge that we possess. Animals have a different sort of knowledge because they 'do not judge of their own judgements, but follow the judgement imprinted on them by God'. Human beings are able to judge their own actions through their powers of reason and their choices are guided by an understanding of the means to an end and the anticipation of that end. Unlike animals, human beings are the cause of their own judgements and actions.

"...the *intellectus agens,* the mind's concept forming power, is likened to a light that enables the mind's eye to see the intelligible features of things, as the bodily eye sees colours"; "...when we frame a judgement in words, our use of concepts is compared, not to seeing something, but rather to forming a visual image of something we are not now seeing, or even never have seen."; "...it is a main thesis in Aquinas's theory of knowledge that what our understanding grasps primarily and most readily is the specific nature (*quod*

quid est) of material substances, in spite of his holding that the senses are in no way cognizant of this nature," - P.T. Geach, Mental Acts, 2nd Ed. [Derived from the *Summa Theologica* of Aquinas]

John Duns Scotus 1265-1308

Duns Scotus put together a reaction against the combined work of Thomas Aquinas and Augustus. He saw the will as being more important than the intellect in the pursuance of the concept of God. He drew a sharp distinction between faith and reason.

The free-will of the individual is of the most importance to the 'being' of the individual; 'the will commanding the intellect is the superior cause of the action'. The intellect being the cause of the willing is subservient to the intention itself. At this time the will was seen as something that was self-orientated and selfish and Aquinas had tried to overcome this by asserting that the intellect was superior to the will.

However, this was contrary to Christian theory and Scotus attempted to overcome this by suggesting that the will had two ends; the first was for the good or advantage of the self, and the second was for the achievement of a more general justice for all things. In the second instance things are valued for their own sake and not because they benefit the individual. Because the will allows one to do something other than what is solely for the individual's advantage it is in this sense a truly 'free' will.

Appendix 2

Intensional language

Briefly, the argument put forward in the discussion of intensional language, i.e. those that express propositional attitudes like 'belief' and 'suppose', is that it is not possible to substitute one term of a sentence for another whilst maintaining the truth-value of the sentence. W.V. Quine sets out two sorts of belief statement that are

feasible, namely 'transparent' and 'opaque'. The 'transparent' sense is that in which it is possible to substitute a term but it is not possible to say whether the truth-value has been changed. For example, it is possible to say truthfully of John that 'he believes that Hesperus is the Morning Star', but it is not possible to say that 'he believes that Phosphorus is the Morning Star' for although we are aware of an identity relation between Hesperus and Phosphorus we cannot be sure that John is also aware of that relation.

Substitutions that are 'opaque' are those in which the truth-value is altered. On the whole Quine claims that the terms in propositional attitudes statements are not intersubstitutive, salve veritate; which is to say that they are 'referentially opaque'. 'Quantifiers and Propositional Attitudes' in *The Ways of Paradox* (New York: Random House, 1966), pp. 183-94.

Appendix 3

System architecture

For architecture I mean the internal structure of the system and the physical constitution of what houses it. So by 'physical constitution' I mean whether the system is organic or inorganic. Organic systems are those made of flesh and blood, and inorganic systems are those that are made artificially. The latter are otherwise described as 'artifacts'.

A broad physical difference, like the external make-up of different systems, allows us to draw an obvious, superficial distinction. The differences in physical characteristics can become infinitely subtle Having a look at the outside of a system can often give us important information about the structure of the innards.

Complexity classes

There are two sorts of problem to face, those for which there is an algorithmic solution and those for which there is not. It is the former case with which I have most interest because I want to find out what computer resources are needed for their execution. Complexity theory investigates this whole area of computational resources. Of the latter case, non-algorithmic problems, it might be feasible to think of them as states that are, as yet, impossible to implement in anything other than a organic system; and then their implementation is something that is intrinsic to the system rather than something that needs to be instantiated from outside.

The resources that are most important are time, memory and hardware. The general term 'time' used to describe the period it takes to execute an algorithm, and 'memory' is the amount of storage required for the algorithm. Memory becomes necessary if partial results are required again later in the execution of the algorithm so that the old computation does not have to be worked over again. The 'hardware' element refers to the amount of actual physical mechanism (e.g. the processor) that is needed for the successful running of the program. In sequential machines it has been found that memory can be traded off for processing time.

Algorithms are designed to accept relevant input data and process it, so the resources that the algorithm needs will vary with the size of the input data. It is possible that different algorithms, using different levels of resources, can be used to solve the same problem. It is probably best and most interesting to use the algorithm that needs fewest resources. Quite often it happens that when one resource is reduced another may have to be increased. Again this is a 'trade-off' situation and the choice of resources must suit the specific application.

Asymptotic Behaviour

The amount of resource that an algorithm uses depends upon the size of the input data. For example, the more digits there are in a calculation the greater the time taken to perform the whole calculation. Sometimes with an increase in the number of digits a term in the function that expresses the amount of resources may begin to dominate other terms. This sort of action is called the *asymptotic* behaviour of the algorithm. Ultimately it is this behaviour which governs the feasibility of a particular algorithm.

With the execution time at n the execution time of the algorithm is proportional to the number of characters of input data. It is easy to see from this that twice as much input data will mean twice as much running time. This amount of time is generally required because the algorithm must, at least, scan the data, (unless, of course, the problem is very trivial). Running on $\log n$ is only possible on a parallel computer (see below) because it can examine many parts of the data simultaneously.

Exponential and Polynomial Algorithms

Related to asymptotic behaviour we have *exponential* algorithms. These have asymptotic behaviour of c^n where c is a constant. These kind of algorithms are not of much use unless the size of the input data is very small. The other kind of algorithms are *Polynomial* and are those where the behaviour is n^c. They are feasible for most, but not all, practical input sizes.

It appears that at a first glance the only feasible algorithms are those that can be executed in a polynomial amount of time. Complexity theory tries to make more clear the distinction between feasible and unfeasible algorithms. Certainly we can expect that certain properties will obtain for the operation of feasible algorithms, and this seems to be the case for those with polynomial time (sequential) algorithms. If we were to combine two feasible algorithms the new algorithm will, predictably, be feasible. These sorts of properties are called *closure* properties and they hold for polynomial algorithms.

To estimate exactly the amount of time that an algorithm will take it is necessary to know something about the internal structure of the computer that it is being run on. All sequential computers have related execution times, which means that each can simulate the other without any significant time loss. So any polynomial algorithm that can be run on one (sequential) computer can also be run on any other (sequential) computer. So it is reasonable to talk of polynomial algorithms independently of any specific computer. This is the *sequential computation thesis;* claiming that all feasibly computable problems are the same for all computers. (It holds for all computers known to date.)

The amount of resource that an algorithm uses is expressed as a function of the input size. But for a given input size there are any number of different inputs; in these cases different algorithms may well use different amounts of resource to deal with the varying types of input data. It is possible for an algorithm to test the input data and perform certain actions that depend upon the outcome of the test.

Worst Case, Average Case and Standard Deviation

As mentioned different amounts of resource are needed for different functions and the amount of resource depends upon the amount of input data. In some cases it is vitally important to know the largest amount of resource that an algorithm might use on a particular given input size, ie to be aware of the longest time it will take a computer to respond in a certain circumstance. This is called the *worst-case* complexity of the algorithm. There is also *average-case* complexity where it is best to know the average that is used over all the inputs of a given size. Finally there is *standard deviation* when the knowledge required is what are the chances of an algorithm remaining close to the average behaviour of a given input.

Upper and Lower Bound

All of the above have been considerations of algorithm complexity in relation to some specified resource. It is also worthwhile talking about the complexity of problems

in relation to the resource. In this instance we are concerned with the complexity of the *best* algorithm being used to solve a particular problem. Being the best algorithm is not always as easy a matter to resolve as it may at first seem since in many cases we are only dealing with the best algorithm so far discovered. This is called the *upper bound* on the complexity of a problem. In some cases it is possible to show that there is a *lower bound* on the amount of resources that an algorithm uses to solve a certain problem. The better the algorithm the lower the upper bound.

Recurrence Relation

A good way of devising the best possible algorithm for solving a problem is to divide it into smaller and smaller problems. This leaves only the smaller problems to be solved. The algorithm sorts through the first half of the problems and then sorts through the second half, the two halves are combined in time n proportional to a constant c mentioned earlier. This procedure is called a *recurrence relation;* which is a sort of divide and conquer technique. The solution of such a procedure will express the resource usage in a very clear way.

NP-completeness and NP-hard

With arbitrary value inputs for a problem there seems to be no straightforward or for that matter, quick method of finding a solution. However, once a method is discovered it is relatively easy to check that the it is right. It can be seen to be the case that for every problem there is an algorithm that can be used for verification in polynomial time as long as we have a proposed set of values and a proposed solution. Overall the solution is the most difficult aspect of the problem to overcome but when found it appears to be obvious.

Problems with a very fast algorithm are called NP. All feasible problems will be in NP since it is possible to both verify the solution of a feasible problem in polynomial time, and to find the solution in the first place in polynomial time. However, the set of

NP is not interesting solely for this reason, it also contains a great many open problems yet to be solved. Open problems are among the hardest because if it was possible to find an algorithm for one it would be possible to find an algorithm for all. An NP problem in the 'open problem' category is called *NP-complete.* Following from this it would seem that they are all computationally equivalent because any polynomial algorithm devised to solve one could be used to solve all the others. If it can be shown that any one of these NP problems is unfeasible it will be an NP-complete problem and this will mean that all NP-complete problems are unfeasible.

All feasible problems have fast algorithms and it is widely accepted that NP-complete problems do not have fast algorithms so the tendency seems to be to believe that they are unfeasible. Any problem that can be reduced to an NP-complete problem is described as NP-hard since it is going to be at least as hard as they are. An NP-hard problem can only be solved in polynomial time if P = NP, where 'P' is equal to the set of problems that can be solved using a temporally bound Turing Machine.

Parallel Computers

In the main complexity theory is concerned with sequential time and memory as the two most important resources. There are now two new, and equally, important resources that have been introduced with the advent of parallel computers, they are "parallel time" and the number of processors needed for the successful execution of the algorithm. 'Parallel time is the time taken to execute an algorithm by a number of processors operating in unison.' (A sequential computer has only a single processor.)

One of the distinguishing features of parallel computers is that they contain a huge number of processors, (perhaps as many as one million), which is similar to the memory contained in sequential computers. All of these processors work at once on one aspect of the problem in the larger algorithm. This means that the algorithm can be

executed in a much shorter space of time; maybe even in real time in a manner not dissimilar to the activity of the brain.

Synchronicity

Another distinction that can be made between parallel and sequential computers is that parallel computers act *synchronously,* ie they perform their computations in unison. Because they are working at the same time communicating between computers is generally made easier.

The parallel computation thesis states that all parallel computer designs are related in their computational abilities. Which is to say that given a particular number of processors an algorithm can be simulated on another parallel computer in roughly the same amount of time. So parallel computations can happily be considered independently of the computer. Of course, this cannot be considered to be a distinction between them and serial computers since their computations can also be considered independently of the computer.

As far as the relationship between memory resources and time used for a computation is concerned, a small amount of memory in a serial computation is proportional to the amount of time it would take on a parallel computer.

It has been mentioned a couple of times that computations on a parallel computer take less time, but this is not always the case. It will always depend on the computation that is being executed. Some tasks are still more suitable for serial computers and would as a result take more time to do on a parallel computer, if they could be done at all. The question to ask is can the task be divided into subtasks that are relevantly independent and if it is then it is suitable for parallel distributing system. If the computation calls for a great deal of communication between the smaller and smaller machines then it will take a lot more time on a parallel computer. The optimal status for an algorithm on a parallel computer is for a computation to have high localised work

that can be done independently of other parts of the distributed net thus calling for low communication between the parts that are doing the individual computations.

Financial Outlay

It has been assumed that not only do more processors mean less time but that they also mean more cost, again this is not always true. More processors can often mean less expense if we are talking in terms of the power of the processors. Something with one processor that is very powerful is going to cost a lot more than a parallel processor with a hundred processors that are weaker. In fact a parallel processor can mean less initial financial outlay with more processing power than one large shared vax.

To execute a program a sequential computer needs time proportional to n to enable it to add together n numbers. A parallel processor can execute the algorithm in time proportional to log n but it needs n processors. For the quickest sequential algorithms time proportional to $n\log n$ is needed.

Appendix 4

Propositions that express belief

When we talk of propositional attitude statements the examples that spring to mind are usually of belief statements. There is a special category of problem inherent in propositions that express beliefs and I will look briefly at how Dennett deals with it. Then I will move on to examine Dennett's proposed answer to the difficulties we encounter overall when using propositional attitude statements. And, in keeping with his recommendation to adopt the intentional stance, we will see that he agrees with the continued use of propositional attitude statements both to describe and predict the behaviour of others.

The difference between *de re* and *de dicto* beliefs is that the former are literally beliefs that are held of or about something (which can be an object or a state of affairs),

and the latter are beliefs that are held of or about a proposition. De re beliefs are usually more specific than de dicto beliefs. For instance, and staying with Dennett's (now outdated!) example (p.168-169), "Bill believes *that* the captain of the Soviet Ice Hockey team is a man". This type of belief is de dicto because it is not held first hand about the thing, but instead is about the proposition itself, and it is the 'that' relation which is crucial to our understanding of the proposition as a whole. "Bill believes *of* his own father that he is a man"; in this proposition the 'of' relation is important, and since Bill has first hand knowledge of his father, that is, of the object in the proposition, the belief is de re.

Dennett disagrees with the maintenance of the de re/de dicto distinction. The distinction has taken other forms, examples are, *relational* and *notional*, and *general* and *specific*. If it is possible to make these distinctions clear then it is also possible to see that the criteria for *de re* belief are very loose. I do not intend to go into this in great detail because it is not directly relevant to the thesis; suffice to say that Dennett concludes that although it is possible to isolate a subset of beliefs that fit the *de re* criteria they are of no theoretical interest to psychology.

Dennett concludes that we have to abandon 'Russell's' Principle: *It is not possible to make a judgment about an object without knowing what object you are making a judgment about.* This will enable us to clarify a number of other linguistic distinctions which have previously been shadowed by the importance that the *de re/de dicto* distinction has assumed. Two of the four conclusions in the "Reflections" section are firstly, that no stable distinction exists between *de re* and *de dicto* beliefs, and secondly, that the Russellian principle ought to be abandoned in the hope of opening up areas for fresh enquiry.

Appendix 5

Is ascription just the over-extension of metaphors?

The ascription of mental states and intentionality to non-human systems has been criticised on a number of fronts, and not least of all by those who say that ascription is simply an over-extension of the metaphorical use of words. I would like to discuss this criticism now before moving on to give a summary and conclusory note of the main points that have been made in this chapter.

A metaphor is used as a figure of speech when we want to talk about something as being that which it only resembles. In this sense, then, the use of human intentional terminology to describe the behaviour of non-human systems is justified for it only suggests that the behaviour resembles that of a human system and not that they are both identical. It is, as Searle say, a "simulation" of human behaviour and not a "duplication".

Other examples of simulation and duplication, or real and artificial, are diamonds, rubber, works of art, and so on. Artificial things become difficult to accept as artificial when they seem identical in nature to the real thing, for how then can any distinction be made between the real and the fake? If we look at all the physical properties of an artificial diamond, that has been made in a laboratory, and discover that essentially it is no different from the real thing that has been mined then the only way to maintain a distinction is to look again at the properties and see if anything has been overlooked or not included. One way to maintain a distinction is to include in the list of properties, the genetic or historic criteria, for how a thing came about. Thus, the distinction would be that one diamond had been made in a laboratory and the other has been created naturally. Their respective genetics now make them distinct.

For an organic system it might well come down to their history being the only property that will eventually make their mentality absolutely distinct from the mentality

exhibited by inorganic systems. Even now we hear functionalists arguing that the human mentality is only just a function of the requisite mental states and no more, so that every aspect of human mentality is, in principle, programmable and it is only a matter of time before we have a fully functioning artificial brain. If that day ever comes the carbonists will have to plead history as a property of mentality and draw their distinction from there.

It is fairly accurate to speak of the metaphorical use of descriptive terms for mental states for with metaphors the comparison between two things is implied rather than explicitly stated. But when we say of a thermostat that it 'knows' when the temperature has dropped we are using the simile 'as-though' which if read in full would be: 'the thermostat has switched the central heating on as-though it knows that the temperature has dropped'. To use 'as' and 'like' is indicative of the use of simile but because of the compression and implicit nature of the comparison between human and non-human mentality we must be dealing with genuine and accurately used metaphors.

That it might be an over-extension of the metaphor would only be possible if the use of such language is restricted to organic systems of the human type, and as we have seen this is not so because we have for as long as we care to remember used the same language with much perceived success to describe the behaviour of non-human animals. We use mentalistic language as an interactive tool with all manner of non-human systems for it makes it possible to make sense of our environment; and within the comparatively recent context of AI it allows us to grasp new inter-disciplinary concepts without having to create and learn a whole new lexicon.

Appendix 6

The points and plots for a two dimensional diagram using *Mathematica*

```
LabelText[String_, Point_] := {}

LabelText[String_, Point_] :=
```

```
Text[FontForm[String, "Italic", 7], Point, {-1, 0}]
  SusansPoints = Table[{LabelText["Human Being", {1.0, 1.2}],
                        Point[{1.2, 1.2}],
                        LabelText["Primate", {1.05, 1.06}],
                        Point[{1.18, 1.15}],
                        Point[{1.12, 1.17}],
                        Point[{1.12, 1.12}],
                        Point[{1.09, 1.12}],
                        Point[{1.15, 1.05}],
                        LabelText["Mammal", {0.97, 0.97}],
                        Point[{1.11, 0.86}],
                        Point[{0.99, 1.0}],
                        Point[{1.08, 0.95}],
                        Point[{1.03, 0.98}],
                        Point[{1.07, 0.9}],
                        Point[{1.0, 0.94}],
                        Point[{1.02, 0.88}],
                        Point[{0.99, 0.91}],
                        Point[{0.93, 0.90}],
                        Point[{0.94, 0.97}],
                        LabelText["Reptile", {0.8, 0.78}],
                        Point[{0.86, 0.74}],
                        Point[{0.81, 0.81}],
                        Point[{0.79, 0.81}],
                        Point[{0.8, 0.77}],
                        Point[{0.74, 0.78}],
                        Point[{0.78, 0.70}],
                        Point[{0.81, 0.71}],
                        LabelText["Bird", {0.7, 0.6}],
                        Point[{0.8, 0.58}],
                        Point[{0.74, 0.54}],
                        Point[{0.705, 0.54}],
                        Point[{0.66, 0.57}],
                        Point[{0.69, 0.59}],
                        Point[{0.67, 0.57}],
                        Point[{0.64, 0.62}],
                        LabelText["Insects", {0.56, 0.47}],
                        Point[{0.62, 0.42}],
                        Point[{0.61, 0.50}],
                        Point[{0.56, 0.55}],
                        Point[{0.56, 0.50}],
                        Point[{0.53, 0.50}],
                        Point[{0.57, 0.39}],
                        Point[{0.53, 0.41}],
                        Point[{0.52, 0.46}],
                        LabelText["Vertebrates", {0.4, 0.26}],
                        Point[{0.43, 0.24}],
                        Point[{0.41, 0.29}],
                        Point[{0.39, 0.23}],
                        Point[{0.37, 0.23}],
                        Point[{0.37, 0.26}],
                        LabelText["Invertebrates", {0.24, 0.1}],
                        Point[{0.28, 0.095}],
                        Point[{0.27, 0.115}],
                        Point[{0.21, 0.11}],
                        Point[{0.195, 0.11}],
                        Point[{0.21, 0.16}],
                        LabelText["PDP's", {0.12, 0.22}],
```

```
                    Point[{0.205, 0.23}],
                    Point[{0.25, 0.26}],
          LabelText["Amoeba, Thermostats etc.", {0.04, 0.04}],
                    Point[{0.09, 0.075}],
                    Point[{0.03, 0.075}],
                    Point[{0.01, 0.06}]}]
(Text[FontForm[Human Being, Italic, 7], {1., 1.2}, {-1, 0}],

Point[{1.2, 1.2}], Text[FontForm[Primate, Italic, 7], {1.05, 1.06},

{-1, 0}], Point[{1.18, 1.15}], Point[{1.12, 1.17}], Point[{1.12,

1.12}], Point[{1.09, 1.12}], Point[{1.15, 1.05}],

Text[FontForm[Mammal, Italic, 7], {0.97, 0.97}, {-1, 0}],

Point[{1.11, 0.86}], Point[{0.99, 1.}], Point[{1.08, 0.95}],

Point[{1.03, 0.98}], Point[{1.07, 0.9}], Point[{1., 0.94}],

Point[{1.02, 0.88}], Point[{0.99, 0.91}], Point[{0.93, 0.9}],

Point[{0.94, 0.97}], Text[FontForm[Reptile, Italic, 7], {0.8, 0.78},

{-1, 0}], Point[{0.86, 0.74}], Point[{0.81, 0.81}], Point[{0.79,

0.81}], Point[{0.8, 0.77}], Point[{0.74, 0.78}], Point[{0.78, 0.7}],

Point[{0.81, 0.71}], Text[FontForm[Bird, Italic, 7], {0.7, 0.6}, {-1,

0}], Point[{0.8, 0.58}], Point[{0.74, 0.54}], Point[{0.705, 0.54}],

Point[{0.66, 0.57}], Point[{0.69, 0.59}], Point[{0.67, 0.57}],

Point[{0.64, 0.62}], Text[FontForm[Insects, Italic, 7], {0.56, 0.47},

{-1, 0}], Point[{0.62, 0.42}], Point[{0.61, 0.5}], Point[{0.56,

0.55}], Point[{0.56, 0.5}], Point[{0.53, 0.5}], Point[{0.57, 0.39}],

Point[{0.53, 0.41}], Point[{0.52, 0.46}], Text[FontForm[Vertebrates,

Italic, 7], {0.4, 0.26}, {-1, 0}], Point[{0.43, 0.24}], Point[{0.41,

0.29}], Point[{0.39, 0.23}], Point[{0.37, 0.23}], Point[{0.37,

0.26}], Text[FontForm[Invertebrates, Italic, 7], {0.24, 0.1}, {-1,

0}], Point[{0.28, 0.095}], Point[{0.27, 0.115}], Point[{0.21, 0.11}],

Point[{0.195, 0.11}], Point[{0.21, 0.16}], Text[FontForm[PDP's,

Italic, 7], {0.12, 0.22}, {-1, 0}], Point[{0.205, 0.23}],

Point[{0.25, 0.26}], Text[FontForm[Amoeba, Thermostats etc., Italic,

7], {0.04, 0.04}, {-1, 0}], Point[{0.09, 0.075}], Point[{0.03,
```

```
0.075}], Point[{0.01, 0.06}]}
Show[Graphics[{SusansPoints}],{Axes->Automatic, AspectRatio >1 }]
-Graphics- etc.
```

REFERENCES

Ambrose, A. (1982) *Wittgenstein's Lectures-Cambridge1932-1935*, (second edition), Oxford, Basil Blackwell.

Barwise, J. & Perry, J. (1983) *Situations and Attitudes*, MIT Press

Bechtel, W. (1992) Studying the Thinking of Non-Human Animals, *Biology and Philosophy*, Vol.7, Pgs.209-215

Beer, C. (1992) Conceptual Issues in Cognitive Ethology, *Advances in the Study of Behaviour*, Vol.21, Pgs.68-105

Bishop, J. (1989) *Natural Agency and essay on the causal theory of action*, Cambridge University Press

Boden, M. (1977) *Artificial Intelligence and Natural Man*, (first edition), Harvester Press

Boden, M. (ed) (1990) *The Philosophy of Artificial Intelligence*, (first edition), Oxford Readings in Philosophy, Oxford University Press

Brand, M. (1984) *Intending and Acting*, (first edition), MIT Press

Casey, G. & Moran, A. (1989) The Computational Metaphor and Cognitive Psychology, *The Irish Journal of Psychology*, Vol.10, No.2, Pgs.143 - 161

Casey, G. (1989) Artificial Intelligence and Wittgenstein, *Philosophical Studies*, Volume XXXII (1988-90), Pgs.156-175

Casey, G. (1990) Minds and Machines, *Royal Irish Academy*

Clark, A. (1988) Thoughts, Sentences and Cognitive Science, *Philosophical Psychology*, Vol.1, No. 3, Pgs.263 - 278

Clark, A. Boden, M. (ed). (1989) *Microcognition - Philosophy, Cognitive Science, and Parallel Distributed Processing*, (first edition), Explorations in Cognitive Science, MIT Press

Cockburn, D. (1991) Human Beings, *The Journal of the Royal Institute of Philosophy*, Supplement: 29, University of Cambridge Press

Davis, W. (1992) The deconstruction of intentionality in archaeology, *Antiquity*, No.66, Pgs.334-347

Dennett, D. C. (1969) *Content and Consciousness*, London: Routledge & Kegan Paul

Dennett, D. C. (1981) *Brainstorms - Philosophical Essays on Mind and Psychology*, (third edition), Harvester Press Ltd.

Dennett, D. C. (1988) Precis of The Intentional Stance, *Behavioral and Brain Sciences*, Vol.11, No.3, Pgs.495 - 546

Dennett, D. C. (1988) *The Intentional Stance*, MIT Press

Deutsch, D. (1992) Quantum computation, *Physics World*, June 1992.

Dretske, F. (1981) *Knowledge and the Flow of Information,* (first edition), Basil Blackwell

Dretske, F. (1983) Precis of Knowledge and the Flow of Information, *Behavioral and Brain Sciences*, Vol.1, No.6, Pgs.55-90

Dretske, F. (1988) *Explaining Behaviour - Reasons in a World of Causes*, MIT Press

Dreyfus, H. (1972) *What Computers Can't Do: A Critique of Artificial Intelligence*, Harper and Row

Fodor, J. (1978) Propositional Attitudes, *Monist*, Vol.61, No.4, Pgs.501-523

Glover, J. (1976) *The Philosophy of Mind*, Oxford Readings in Philosophy, (1980 edition), Oxford University Press

Gordon, D. M. (1992) Wittgenstein and Ant-Watching, *Biology and Philosophy,* Vol.7, Pgs.13-25

Grandy, R. (ed) (1986) *Philosophical Grounds of Rationality*, Grice - Addresses, essays and lectures, Oxford University Press

Grishman, R. (1986) *Computational Linguistics - An Introduction*, Cambridge University Press

Harman, G. (1982) *On Noam Chomsky,* The University of Massachusetts Press

Harnad, S. (1990) The Symbol Grounding Problem, *Physica D* 42 North Holland, Pgs.335-346

Haugeland, J. (1981) *Mind Design*, MIT Press

Hobbes, T. (1651) *Leviathan*, (1984 edition), Penguin English Library

Husserl, E. (1983 - first published 1913) *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy - first book,* The Hague, Martinus Nijhoff Publishers

Jackendoff, R. (1987) *Consciousness and the Computational Mind*, Explorations in Cognitive Science - published in cooperation with The British Psychological Society, MIT Press

Kamppinen, S. S. &. M. (1990) *A Historical Introduction to Phenomenology*, Croom Helm - Methuen

Kant, I. (circa1780) *Critique of Pure Reason*, (second edition 1985), Macmillan

Karmiloff-Smith, A. & Clark, A. (1990 - May) The Cognizer's Innards, *Brain and Behavioural Science*

Kelly, K. (1992) Deep Evolution - The Emergence of Postdarwinism, *Whole Earth Review*, Fall, Pgs.4-21

Kiss, G. &. Han, R. (1991) Towards a Semantics of Desire, Technical Report 70, HCRL, Open University, 30th April 1991

Kiss, G. (1988) Some Aspects of Agent Theory, Technical Report 43, HCRL, Open University, December 1988

Kiss, G. (1991) Variable Coupling of Agents to their Environment: Combining Situated and Symbolic Automata, HCRL, Open University, 1991

Krishnamurthy, E. V. (1983) *Introductory Threory of Computer Science*, Macmillan Press

Langton, C. G. (1984) Self-Reproduction in Cellular Automata, *Physica D* , Vol.10, Pgs.135-144 North-Holland

Langton, C. G. (1989) Artificial Life, *Santa Fe Institute Studies in the Science of Complexity*, Santa Fe, Addison - Wesley

Lee, D. (1980) *Wittgenstein's Lectures-Cambridge 1930-1932*, Basil Blackwell

Lister, G. (1982) *Computer Science a Modern Introduction*, Prentice-Hall International

Looren De Jong, H. (1991) Intentionality and the Ecological Approach, *Journal for the Theory of Social Behaviour*, Vol.21, No.1

Lycan, W. G. (1987) *Consciousness*, MIT Press

Lyons, W. (1990) Intentionality and modern philosophical psychology, I. The modern reduction of intentionality, *Philosophical Psychology*, Vol.3, No.2, Pgs.247 - 269

Lyons, W. (1991) Intentionality and modern philosophical psychology - II. The return to representation, *Philosophical Psychology*, Vol.4, No.1, Pgs.83 - 102

Lyons, W. (1991) Introspection - A two-level or one-level account?: A response to Howe, *New Ideas in Psychology*, Vol.9, No.1, Pgs.51 - 55

Marras, A. (1972) *Intentionality, Mind, and Language*, University of Illinois Press

Maudlin, T. (1989) Computation and Consciousness, *The Journal of Philosophy*, Pgs.407 - 432

Morris, W. E. (1990) Knowledge and the Regularity Theory of Information, *Synthese*, 82, Pgs.375-398, Kluwer Academic Publishers

Natsoulas, T. (1983) Concepts of Consciouness, *The Journal of Mind and Behavior*, Vol.4, No.1, Pgs.13-59

Natsoulas, T. (1991) The Concept of Consciousness, *Journal for the Theory of Social Behaviour*, Vol.21, No.1

Noble, E. (1989) Goals, No-Goals and Own Goals - *A Debate On Goal-Directed and Intentional Behaviour*, Chapter 10 - "Narrow Intentions" - Lockery, S. Oxford University Press

Penrose, R. (1989) *The Emperor's New Mind - Concerning Computers, Minds, and The Laws of Physics*, Oxford University Press

Putnam, H. (1988) *Representations and Reality*, MIT Press

Quine, W.V. (1960) *Word and Object*, MIT Press

Ringle, M. (1979) *Philosophical Perspectives in Artificial Intelligence*, Harvester Press Limited

Rosenschein, S. J. (1985) Formal Theories of Knowledge in AI and Robotics, Technical Note 362, SRI International, 10th September

Rosenschein, S. J. (1987) The Synthesis of Digital Machines with Provable Epistemic Properties, SRI International, Techincal Note 37

Ryle, G. (1949) *The Concept of Mind,* Peregrine Books

Searle (1983) *Intentionality*, Cambridge University Press

Searle, J. (1980) Minds, brains and programs, *Behavioral and Brain Sciences*, Vol.3, Pgs.417 - 457

Searle, J. (1984) *Minds, Brains and Science*, Pelican Books

Searle, J. R. (1990) Consciousness, explanatory inversion, and cognitive science, *Behavioral and Brain Sciences*, No.13, Pgs. 585-642

Shakespeare, W. (circa.1601) *Hamlet,* The New Penguin (1980)

Sloman, A. (1991) AI, Neural Networks, Neurobiology, Architectures and Design Space, *AISB Quarterly Newsletter of the Society for the Study of Artificial Intelligence and Simulation of Behaviour*, No.78, Pgs.10-13

Sloman, A. (1991) Prolegomena to a theory of communication and affect, *NATO Advanced Research Workshop*, "Computational theories of communication and their applications: Problems and Prospects", No.24, Edited by Ortony, A. & Slack, J., Springer Verlag

Stich, S. P. (1975) *Innate Ideas,* University of California Press

Stuart, S. A. J. (1991) Mental Properties of Systems, Technical Report 68, HCRL, Open University

Stuart, S. A. J. (1992) Mental States and Intentionality: A Review of the Literature, Technical Report 76, HCRL, Open University

Stuart, S. A. J. (1992) When is it justifiable to ascribe mental states to non-human systems?, Forthcoming publication in the *Proceedings of the Mind and Related Matters Conference*, Leeds, September 1992

Stutt, A. (1989) Argument in the humanities: a knowledge-based approach, PhD Thesis, Human Cognition Research Laboratory

Tighlman, B. R. (1991 - July) What is it Like to be an Aardvark? *The Journal of the Royal Institute of Philosophy*, Cambridge University Press, Vol.66, No.257, Pgs.325-338

Torrance, S. (1984) *The Mind and The Machine*, Ellis Horwood Limited

Watt, S. (1992) Labyrinths, Chaos, and the Fractal Geometry of the Mind, Technical Report, HCRL, Open University

White, A. R. (1968) *The Philosophy of Action*, Oxford Readings in Philosophy, Oxford University Press

Wilkes, K. (1978) *Physicalism*, Routledge and Kegan Paul

Wittgenstein, L. (1958) *Philosophical Investigations*, Basil Blackwell
Ziff, P. (1966) *Philosophic Turnings - Essays in Conceptual Appreciation*,